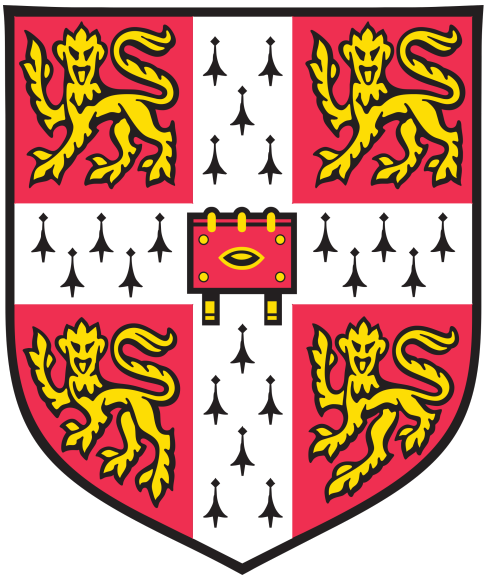
Investigating 3D genome organisation in *Caenorhabditis elegans* with Accessible Region Conformation Capture (ARC-C)



This dissertation is submitted for the degree of Doctor of Philosophy in Genetics

Wei Qiang Seow

Churchill College

September, 2018

DECLARATION

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

It does not exceed the prescribed word limit for the relevant Degree Committee.

ABSTRACT

Wei Qiang Seow

Investigating 3D genome organisation in Caenorhabditis elegans with Accessible Region Conformation Capture (ARC-C)

3C and its derivatives have been applied in various organisms to study chromatin architecture. However, these methods have limitations: most of them are limited to restriction-fragment resolution and all of them, with the exception of Hi-C, only survey a pre-defined subset of the genome. I developed a variant of Hi-C, named Accessible Region Conformation Capture (ARC-C), in *C. elegans*, which interrogates genome organisation at multiple scales genome-wide - from domains and compartments to high resolution interactions between regulatory elements. I applied ARC-C in wild-type Bristol N2, *met-2 set-25* mutants that have no H3K9 methylation to study the effects on domain and compartment formation. In these mutants, compartmentalisation (i.e. inter-domain interactions) between H3K27me3-enriched regulated domains is reduced. I also used ARC-C in *blmp-1* mutants to understand the role of BLMP-1 in chromatin looping. In *blmp-1* mutants, interactions between putative BLMP-1 mediated loops for downregulated genes are significantly reduced. In wild-type worms, when surveying significant interactions at 500 bp resolution, I observe the presence of dense clusters of significant interactions anchored at high occupancy target (HOT) regions that I call "hubs." Interestingly, the deletion of these hubs does not affect the transcription of linked or local genes. However, local interactions are altered and some extent of redundancy is observed. To improve our ARC-C protocol, I tested several variations with different enzymes and biotin-mediated streptavidin beads pulldown. In all, ARC-C revealed insights into genome organisation in *C. elegans* and I have made progress toward a next-generation version of the method.

ACKNOWLEDGEMENTS

I am not particularly effusive when it comes to acknowledgements and compliments, a trait that was perhaps reinforced by my time in the UK, but the arduous journey to finish up this thesis has weakened my defences considerably. First, I would like to thank my supervisor, Professor Julie Ahringer, for the opportunity to work in her laboratory. I appreciate her faith and guidance when I was thrusted into a burgeoning field at a time when people were still wrangling over their interpretation of Hi-C data. Her tutelage was invaluable and a character-building experience.

I would also like to thank my scholarship provider, A*STAR, for the opportunity to study in Cambridge and the forbearance to tolerate my tardiness when it comes to annual reports. Thank you for my ability to shop at Waitrose a little more frequently than I would have.

Thank you to the Department of Genetics, and my advisor, Professor Anne Ferguson-Smith, for her incisive mind and helpful critique.

My time in Cambridge has been made much more joyful with the addition of my dear friends Rohan, Gail, and Hanae. Thank you, Rohan and Gail, for inviting me into your home and tutoring me on the idiosyncrasies of British life - I will endeavour to say “banana” the “proper” way henceforth and also use “trousers” instead of “pants”. To Rohan and Hanae, I still think Suicide Forest is an apt name for a band, but am agreeable to Various Artists now.

My family deserves great affection and credit for being incredibly supportive, despite their unwillingness to read beyond two sentences in my thesis.

To Garima, thank you for making my days in the laboratory less tedious and more enjoyable. It would not have been the same without you.

I would like to thank my fellow students - special mention to Carson, who left us too soon for the US - for the good times and support during the bad. I think we can all agree that the Christmas party that we organised was the best Gurdon party ever.

To all my friends back in Singapore, thank you for dropping by whenever you had the chance and bringing a bit of home with you.

Finally, I would like to thank Netflix and Spotify for being the best and worst friend one can have during thesis-writing.

TABLE OF CONTENTS

ABSTRACT

ACKNOWLEDGEMENTS

INTRODUCTION 1

I ARC-C DEVELOPMENT AND EVALUATION 33

Principles underlying ARC-C 33

ARC-C 38

Experimental steps 38

Data processing 39

Evaluating ARC-C data 42

Comparison and evaluation with published Hi-C 50

Calling significant interactions 61

Biases 61

Existing methods for correcting biases 62

Experimental setup 63

Results 68

II USING ARC-C TO DEFINE CHROMATIN INTERACTIONS AT HIGH RESOLUTION 80

Regulatory interactions 80

Promoter-promoter interactions 81

Promoter-enhancer interactions 88

Interaction hubs 92

Transcriptional effects of hub deletions 96

ARC-C in hub deletions 101

III FACTORS MEDIATING LOOP FORMATION 117

Factor APA 117

Cohesin and condensin 125

Chromatin regulators 135

Transcription factors 138

ARC-C in *blmp-1* mutants 139

IV DOMAINS AND COMPARTMENTS 147

Domains and compartments in wild-type worms 148

Role of H3K9 methylation in domains and compartments 153

Domains and Compartments in *met-2 set-25* mutants 157

VI NEXT-GENERATION ARC-C 172

Optimisation strategies 172

Next-generation ARC-C 176

Observations 179

In-vitro ngARC-C optimisation 185

Discussion 190

DISCUSSION 192

METHODS 198

REFERENCES 212

INDEX OF FIGURES 236

LIST OF ABBREVIATIONS 244

APPENDIX 247

A1 Supplementary Data 247

INTRODUCTION

Chromosomal DNA is folded and packaged at different hierarchical levels. Practically, this is required to fit approximately 2 m of DNA into a nucleus 2 to 10 um wide in mammals; however, DNA typically only occupies up less than 3% of nuclear volume (Webster *et al* 2009). Functionally, DNA is differentially organised within the nucleus as a cause or consequence of biological function. For example, electron dense heterochromatin is usually found at the nuclear periphery and active, electron-lucent euchromatin in the nuclear interior (Heitz 1928, Fawcett 1966). Chromosomal DNA is separated into discrete chromosomes which occupy their own nonrandom distinct area in the nucleus. Compartments that interact amongst themselves and highly self-interacting domains called topologically associating domains were found in mammals and certain invertebrates (Lieberman-Aiden *et al* 2009, Dixon *et al* 2012, Sexton *et al* 2012). Within these domains, chromatin establishes structural and regulatory loops. Such looping is necessary for coordinating gene expression by bringing distant enhancers into proximity with promoters and can be misregulated in diseases (reviewed in Norton & Phillips-Cremins 2017, Mishra & Hawkins 2017). It is therefore important that we understand the structural and regulatory landscape of chromatin.

Chromatin folding hierarchy

The study of 3D genome organisation and chromatin folding has been aided by advances in microscopy and the chromosome conformation capture (3C) techniques. By microscopy, nuclear localisation and the extent and pattern of chromatin folding can be examined. 3C and its variants rely on the same principles: chromatin architecture is preserved through fixation; thereafter, DNA digestion and proximity ligation create chimeric fragments that confer information on spatial proximity between genomic loci. 3C variants can be categorised according to the number of loci they interrogate (reviewed in Denker & de Laat 2016). 3C analyses interactions between selected pairs of sequences (one-to-one) (Dekker *et al* 2002), circular chromosome conformation capture (4C) identifies all regions interacting with a selected viewpoint (one-to-all) (Simonis *et al* 2006, Zhao *et al* 2006), chromosome conformation capture carbon copy (5C) assays multiple interactions in parallel (many-to-many) (Dostie *et al* 2006), and Hi-C interrogates interactions genome-wide (all-to-all) (Lieberman-Aiden *et al* 2009).

Chromosome territories, nuclear bodies, and trans interactions

Chromosome territories (CT) describe the phenomenon where chromosomes occupy a nonrandom distinct space within the nucleus (Croft *et al* 1999, Cremer *et al* 2006). It was first described in horse roundworm where visible chromosomes retain their discreteness and unique nuclear position during interphase and with

minor movements through cell division (Rabl 1885, Boveri 1909). This phenomenon has been independently corroborated by UV irradiation and pulse labelling (Cremer *et al* 1982), fluorescence *in situ* hybridisation (FISH) (Pinkel *et al* 1988, Bolzer *et al* 2005), and specifically Hi-C studies where intrachromosomal interactions are much more prevalent than interchromosomal interactions, even at thousands of Mb apart (Lieberman-Aiden *et al* 2009). CTs have nonrandom radial distribution (position relative to nuclear periphery and core) which are associated with their gene density (Boyle *et al* 2001, Neusser *et al* 2007) - for instance, the gene poor human chromosome 18 is consistently located at the nuclear periphery while the gene rich chromosome 19 is frequently observed in the nuclear interior (Croft *et al* 1999) - as well as other parameters like chromosome size, transcriptional activity, GC content, and replication timing (Goetze *et al* 2007, Grasser *et al* 2008, Hepperger *et al* 2008). By contrast, there is much less evidence for a nonrandom pattern of proximity (position relative to other chromosomes) between CTs. While chromosomes have positional preference, proximity patterns are stochastic and vary from cell-to-cell; homologous chromosomes can occupy disparate locations with different CT neighbours (Meaburn & Misteli 2007).

Neighbouring CTs intermingle at their borders to coordinate transcription (Branco & Pombo 2006, Belyaeva *et al* 2017). Advances in FISH showed that CTs constitute a condensed core territory and a surrounding "corona" (Boyle *et al* 2011,

Kalhor *et al* 2012) of chromatin looping out of the core into neighbouring CTs, also called "chromosome kissing" (Cavalli 2007). Looped out regions are typically enriched for hallmarks of active chromatin - high transcriptional activity, active histone marks, high density of DNase I hypersensitive sites, and high gene density (Simonis *et al* 2006, Hou *et al* 2012, Sexton *et al* 2012). While looped out regions are gene dense and highly active, such as the 11p15.5 region (Küpper *et al* 2007), major histocompatibility complex (Volpi *et al* 2000, Branco & Pombo 2006), and epidermal differentiation complex loci (Williams *et al* 2002), the reorganisation of genes into the interchromosomal domain does not precipitate transcriptional upregulation (Morey *et al* 2009). Intermingling of CT borders likely aid in transcriptional cooperation and modulation by bringing related genes on different chromosomes into proximity. The *Hoxb* clusters loop out of the core CT region upon activation by cellular cues during gastrulation (Chambeyron *et al* 2005), the α-globin gene cluster interacts with active erythroid genes on other chromosomes (Brown *et al* 2008), mammalian X-chromosomes "kiss" in a process involving the *Xic* locus to count and choose the copy of X to be inactivated (Bacher *et al* 2006, Xu *et al* 2006), and the efficiency of imprinting at the *PWS*/*AS* region (chr 15q11-13) and the *IGF2*/*H19* locus (chr 11p15) was found to be correlated with the proximity between the two loci (LaSalle & Lalande 1996).

Nuclear bodies, which are punctate structures identified by microscopy in nuclei, provide a medium for the organisation of interchromosomal interactions. Olfactory receptor (OR) gene clusters are typically silenced by their interchromosomal convergence to olfactory sensory neuron-specific heterochromatic nuclear bodies (Clowney *et al* 2012). Recent Hi-C and microscopy studies found that the activation of an olfactory receptor (OR) allele out of ~ 2,800 possible options (one neuron-one receptor rule) involves the escape of the chosen OR from these heterochromatic foci to a *trans*-interacting network of euchromatic enhancers or "Greek Islands" (Lomvardas *et al* 2006, Clowney *et al* 2012, Markenscoff-Papadimitriou *et al* 2014, Monahan *et al* 2017).

Separately, two major hubs of non-overlapping interchromosomal interactions were found by a technique capable of mapping multiple and longer range interactions and characterised as either active or inactive (Quinodoz *et al* 2018). The active hub organises around nuclear speckles that are enriched for RNA polymerase II (RNA pol II) occupancy, mRNA splicing and processing proteins, while the inactive hub is closely associated with the nucleolus (Quinodoz *et al* 2018). As suggested from earlier studies where the nucleolar periphery appears to act as a matrix for attaching centromeric clusters (Padeken *et al* 2013), these nuclear bodies serve as anchors to constrain chromatin folding and influence high-order architecture (Quinodoz *et al* 2018).

3D compartments and domains

Within CTs, Hi-C experiments have shown that chromosomes can be partitioned into multi-Mb sized compartments that are broadly either active (A-type) or inactive (B-type). A/B compartments manifest as a "checkerboard" pattern of enriched contact frequency in both inter- and intrachromosomal contact maps and can be estimated by an eigenvector decomposition of the normalised contact matrix (Lieberman-Aiden *et al* 2009). A/B compartments tend to associate with their own compartment-type. In mammals, A compartments are generally gene dense, have higher transcriptional activity, greater accessibility to DNase I, are enriched for active histone modifications such as H3K36me3 and instances of poised chromatin (Lieberman-Aiden *et al* 2009). Conversely, B compartments are gene-poor, enriched for silencing H3K9me3, have high contact frequency, and a higher propensity to be self-interacting (Lieberman-Aiden *et al* 2009). In addition, B compartments tend to be associated with late replication timing (Dekker *et al* 2013), the nuclear lamina and nuclear periphery localisation (Ryba *et al* 2010).

With greater sequencing depth and higher resolution, A/B compartments were further segregated into multiple subcompartments - A1, A2, B1 to B4 - each of which possess unique chromatin properties: for example, B1 enriches for polycomb group (PcG) proteins, B2 for lamina- and nucleolar-associated domains

(LADs & NADs), and B3 for LADs (Rao *et al* 2014). A2 compartments are close in linear distance to LADs (Robson *et al* 2017), likely representing "unlocked" regions looped from the nuclear periphery and nuclear pore-associated regions (Peric-Hupkes *et al* 2010). The radial positioning preferences of these subcompartments were later verified by super-resolution FISH (Stevens *et al* 2017).

Compartments likely represent the transcriptional status of their constituent genes. Compartments switch between A and B types in accordance with cellular differentiation or cell-type differences, in a manner that is consistent and correlated with gene activity (Takebayashi *et al* 2012, Fraser *et al* 2015, Dixon *et al* 2015). The repression of developmental genes and A-to-B compartment switches can happen simultaneously with nuclear lamina association (Criscione *et al* 2016). In addition, A/B compartments can be accurately predicted with just DNA methylation and assay for transposase-accessible chromatin (ATAC-seq) data (Fortin & Hansen *et al* 2015), further reinforcing their association with chromatin activity and gene expression regulation.

Early studies in mammals employing Hi-C discovered the existence of topologically-associating domains (TADs): blocks of self-interacting domains along the diagonal of the contact map; intra-TAD interactions are substantially

more frequent than inter-TAD interaction (Lieberman-Aiden *et al* 2009, Dixon *et al* 2012). TADs are likely similar to ~100 kb to 1Mb replication foci or chromosomal domains that were first described as basic structural units of CTs (Ma *et al* 1998, Albiez *et al* 2006).; TADs share boundaries with replication-timing domains (Pope *et al* 2014). TAD boundaries are enriched for housekeeping genes, cohesins, and insulators like CTCF in mammals (Lieberman-Aiden *et al* 2009, Dixon *et al* 2012) and various insulator proteins in *Drosophila* (Hou *et al* 2012, Sexton *et al* 2012, Stadler *et al* 2017, Ramírez *et al* 2018). In fact, it has been noted that the strength of contact insulation at TAD boundaries correlates with the co-occupancy of insulators or architectural proteins (van Bortle *et al* 2014).

TADs have a median size of around 800 kb in mammals and around 60kb in invertebrates when they were first reported (Lieberman-Aiden *et al* 2009, Zhang *et al* 2012, Sexton *et al* 2012, Nora *et al* 2012), but this is likely conditional on the depth of sequencing. Sub-TADs, which are finer domains within TADs that preferentially interact, were later described and also had CTCFs enriched at boundaries (Phillips-Cremins *et al* 2013, Sofueva *et al* 2013). Strikingly, sub-TADs displayed more cell-type variation than TADs in general (Zhan *et al* 2017).

In contrast to compartments, TADs are largely cell-type and developmental stage invariant (60-70%) (Dixon *et al* 2012, Dixon *et al* 2015, Schmitt *et al* 2016).

Interestingly, TADs correlate with, but do not rely on, certain features of the epigenome, such as H3K27me3, H3K9me2, or LADs (Nora *et al* 2012). That said, this correlation is much weaker than that between compartments and the epigenome (Falk *et al* 2018). The epigenome varies between cell types but TADs remain intact, suggesting separate systems of regulation. Transcriptionally active TADs are also generally more heterogeneous than inactive TADs (Sofueva *et al* 2013). In all, there is a growing consensus that TADs and compartments are formed independently (Schwarzer *et al* 2017, Rao *et al* 2017, Wutz *et al* 2017).

Separately, across *Mus musculus*, *Canis familiaris*, *Macaca mulatta*, and *Oryctolagus cuniculus*, TADs were found to be evolutionarily conserved in syntenic regions; specific interactions at conserved CTCF sites were also conserved between *C. familiaris* and *M. musculus* (Rudan *et al* 2015). The robust conservation was strongly CTCF-dependent and evolutionary rearrangements were mediated by CTCF motif rearrangements (Rudan *et al* 2015). These findings hint at a modularity that is essential for biological function.

Chromatin loops

Dynamics between regulatory elements, such as promoters, enhancers, silencers, and insulators, play an important role in regulating transcription and genome stability, failing which, developmental defects or disease could ensue.

Classically, enhancers are distal elements to core promoters and are known to activate or augment transcription rate for their target promoters (Maniatis *et al* 1987). Regulatory elements can exist far apart from each other in linear distance: the *cis*-regulatory element ZRS regulates *Shh* from almost 900kb away (Lettice *et al* 2003). Large strides have been made in identifying regulatory elements through accessibility assays, histone modification and transcription-factor binding assays, nucleosome positioning assays, and chromatin state mapping (Mendenhall & Bernstein 2008). Even then, annotations of *cis*-regulatory elements based on these proxies can be unreliable depending on the computational approach used (Zacher *et al* 2017). Studies traditionally assign an annotated regulatory element to the nearest transcriptional start site (TSS), but vertebrate enhancers often skip nearby genes (de Laat & Duboule 2013), with possibly only around 7% of regulatory elements interacting with their nearest gene (Sanyal *et al* 2012).

Given observations that enhancer-promoter interactions (EPI) can act in an orientation-independent manner, occur at variable distances, and often skip neighbouring genes (Sanyal *et al* 2012, Pennacchio *et al* 2013), it is likely that some form of chromatin folding is involved. There is independent evidence that proteins mediate chromatin loops: in-vitro electron microscopy experiments showed that Sp1 could connect two regulatory elements via oligomerization (Li *et al* 1991). Importantly, the artificial tethering of a looping factor LDB1 was able to

significantly activate β-globin transcription in GATA1-null erythroblasts through the formation of a chromatin loop (Deng *et al* 2012). Vertebrate CTCF-bound DNA dimerises *in vitro* (Pant *et al* 2004), which was interpreted as evidence they could mediate loops in a fashion similar to *Drosophila* insulator proteins (Gerasimova *et al* 2000).

The advent of high-throughput C-based technologies advanced the notion of CTCF as a "master weaver" of the genome (Phillips & Corces 2009). At the extensively studied β-globin locus, CTCF-bound DNase I hypersensitive sites (DHS) form an active chromatin hub, bringing a cluster of enhancers called the locus control region closer in proximity to activated globin genes in β-globin expressing erythroid cells (Tolhuis *et al* 2002). CTCF-binding sites are enriched with cohesin, which has been independently shown to correlate with active genes (Misulovin *et al* 2008, Dorsett & Merkenschlager 2013); together CTCF and cohesin have been implicated in chromatin loop formation through loop extrusion, where DNA is threaded through ring-like cohesins to form loops, which is delayed or stabilised by CTCF (Sanborn *et al* 2015, Goloborodko *et al* 2016, Fudenberg *et al* 2016).

Although CTCF and cohesin have been shown to be necessary for the formation of loops at TAD boundaries (Nora *et al* 2017, Schwarzer *et al* 2017, Rao

et al 2017), it is unclear the extent to which they are involved in EPI genome-wide. A study in mouse limb and midbrain development identified two classes of interactions - a tissue-type independent, structural type involving CTCF and cohesins, and a spatiotemporally dynamic type enriched for repressive or active histone marks that was defined as regulatory (Andrey *et al* 2017). The former corresponded to peak foci found at the corners of TADs or insulated neighbourhoods in Hi-C contact maps and was thought to indicate stable, structural anchors (Rao *et al* 2014, Dowen *et al* 2014, Hnisz *et al* 2016).

Regulatory interactions and interaction networks

Regulatory interactions are better captured by methods that trade the complexity of Hi-C for increased resolution at subsetted regions of interest, such as 4C, HiChIP, Capture-Hi-C (CHi-C), and Capture-C (Zhao *et al* 2006, Hughes *et al* 2014, Schoenfelder *et al* 2015a, Mumbach *et al* 2016). Surprisingly, based on studies applying these methods, EPIs display a wide range of dynamism. During limb development, the *HoxD* cluster or *Satb1* gene underwent considerable changes in EPIs in a tissue-specific manner via a dynamic TAD boundary (van de Werken *et al* 2012, Andrey *et al* 2013). A larger scale study of 17 haematopoietic cell types revealed that significant EPIs were highly cell-type or lineage specific (Javierre *et al* 2016). However, in *Drosophila* embryos, 94% of enhancer contacts were stable across various time points and tissues despite changes in their target

gene's activity, with gene activation occurring through the release of paused polymerase (Ghavi-Helm *et al* 2014). Indeed, the *Hox* gene network in mouse embryonic stem cells (mESCs) was enriched for contacts with poised enhancers (H3K4me1, H3K27me3) (Schoenfelder *et al* 2015b). Together, the data indicate that EPIs constitute a complex regulatory system that can modulate transcription through dynamic changes in 3D contacts or the underlying chromatin context.

Principles underlying EPIs and gene regulation are not well-established, with separate instances of enhancers acting additively, redundantly, or synergistically; it is unclear what modulates the behaviour of enhancers. At the α-globin locus in mice, a putative super enhancer was shown to be a cluster of independently acting enhancers; combinatorial deletions of the components in the super enhancer revealed an additive effect and no unexplained synergism for globin gene expression (Hay *et al* 2016). The same conclusions about super enhancers were drawn in a study of the *PIM1* oncogene (Xie *et al* 2017). Similarly, consecutive deletions in a constellation of enhancers that control the Indian hedgehog gene (*Ihh)* indicated that these enhancers were phenotypically redundant but regulate gene expression additively and each enhancer contributes differently in different tissue-types (Will *et al* 2017).

Additivity can be mediated synergistically through enhancer-enhancer interactions; in the case of the mouse *Krox20*, an enhancer was needed to potentiate the activity of another enhancer (Thierion *et al* 2017). Intriguingly, enhancers can display conditional non-additive characteristics. By comparing pairs of primary and shadow enhancers in *Drosophila* embryos, Bothma *et al* (2015) found that weak *knirps* enhancers are additive likely because they do not interact frequently with the *knirps* promoter, while strong *snail* enhancers compete for promoter access and are sub-additive: the deletion of the proximal enhancer unexpectedly increases *snail* expression. At the *hunchback* locus, enhancers display sub-additivity with saturating amounts of Bicoid activator but become additive at lower levels (Bothma *et al* 2015). Put together, while EPIs are generally additive when considered discretely, gene regulation goes beyond mere sequences, with chromatin state, enhancer activity, and synergistic feedback loops all part of the complex interplay of factors.

Genes are organised by transcription factors and other chromatin regulators into higher order spatial interaction networks, presumably for co-regulation. In pluripotent cells, pluripotent genes interact both in *cis* and in *trans* within a common space, sharing general transcriptional machinery and specific factors (de Wit *et al* 2013). Pluripotency factors (like OCT4, SOX2, NANOG, and KLF4) are enriched for long-range interactions. These factors mediate or stabilise the

chromatin structures that bring pluripotency genes together (Bouwman & de Laat 2015, Stevens *et al* 2017): for instance, KLF4-mediated loops were lost upon differentiation or the deletion of *Klf4* (Wei *et al* 2013). Chromatin regulators and architectural proteins such the mediator complex and cohesin are enriched over interacting sites bound by pluripotency factors (Phillips-Cremins *et al* 2013). These interactions are lost upon differentiation, implicating pluripotency factors in the recruitment or tethering of mediator and cohesin (de Wit *et al* 2013). Pluripotentcy factors appear central to the process of maintaining pluripotency and are considered "master regulators" (Rizzino 2009, Davis & Rebay 2017) but it is as yet unclear the extent to which they underlie and influence transcription and chromatin architecture.

Concomitant with pluripotency factor-mediated interaction networks, developmental genes in pluripotent cells are repressed by Polycomb group (PcG) proteins. RING1B, a core component of PRC1, maintains a PRC2-independent central *Hox* network of genes (*Hox* cluster and 66 connected genes), which is abrogated upon *Ring1A*/*Ring1B* double knockout (Schoenfelder *et al* 2015b). Separately, the knockout of *Eed*, which is part of the PRC2 complex, results in the weakening of clustering between Polycomb/H3K27me3 regions (Denholtz *et al* 2013) but not the *Hox* network (Schoenfelder *et al* 2015b), suggesting independent mechanisms of spatial clustering. The clustering of PcG complexes manifests as

nuclear bodies (Pirrotta & Li 2012, Cheutin & Cavalli 2014, Wani *et al* 2016), which likely represents a higher order form of repression through sequestration and chromatin compaction (Boettiger *et al* 2016). Similarly, the nucleosome remodelling deacetylase (NuRD) clusters stochastically with active enhancers and promoters in 3D to form nuclear foci (Stevens *et al* 2017), potentially regulating chromatin activity in a manner similar to interchromosomal active hubs and nuclear speckles. A reductionistic approach to the study of EPIs can be helpful for elucidating general principles, but it has be to combined with an appreciation of regulatory interaction networks for a fuller understanding of chromatin architecture and gene regulation.

Chromatin architecture & disease

Given how crucial genome organisation is to genome stability and gene regulation, it stands to reason that aberrations could lead to diseases. TADs regulate gene expression by delineating the extent of regulatory signals; the deletion of CTCF-bound TAD boundaries resulted in ectopic interactions between regulatory elements from multiple TADs (Dowen *et al* 2014). In patients with T-cell acute lymphoblastic leukemia, the weakening of an insulated neighbourhood activated proto-oncogenes *TAL1* or *LMO2* (Hnisz *et al* 2016). Structural variations (deletions, duplications, or inversions) can cause developmental abnormalities. When re-engineered in mice via CRISPR-Cas9 (Kraft *et al* 2015), structural

alterations can lead to alterations in TAD boundaries and create pathogenic misregulation. In one such case, an inversion that put *Epha4* enhancers under the control of a *Wnt6*-containing TAD and near the *Wnt6* gene resulted in the ectopic expression of *Wnt6* and a digit malformation called F-syndrome (Lupiáñez *et al* 2015). Duplications can produce new chromatin interaction domains (neo-TADs), which drive ectopic expression if they encapsulate mismatched regulatory units. A neo-TAD encompassing a native *Kcnj2* gene in mice and duplicated *Sox9* enhancers led to a *de novo* expression of *Kcnj2* with a *Sox9* spatiotemporal expression pattern, with shorter digits and nail aplasia (Franke *et al* 2016).

Factors that are closely associated with modulating chromatin architecture are mutated or misregulated in many diseases. Patients with heart failure can show diminished CTCF binding; the use of left ventricular assist devices alleviates this issue with an increase in CTCF abundance (Rosa-Garrido *et al* 2017). CTCF-knockout mice models experience cardiomyopathy; genomically, they have weaker boundary insulation, altered A/B compartmentalisation and genome accessibility corresponding to a rewiring of regulatory interactions and the activation of the fetal gene programme (Rosa-Garrido *et al* 2017). Factors that affect CTCF-binding can also be pathogenic. Hypermethylation of CTCF-binding sites leading to reduced CTCF-binding silenced the *XAF1* gene through the loss of

a chromatin loop (Victoria-Acosta *et al* 2015) and caused an overexpression of an oncogene by allowing access to illegitimate enhancers (Flavahan *et al* 2016).

In Hutchinson-Gilford progeria syndrome, a premature ageing disorder, a point mutation in lamin A, which is responsible for tethering chromatin to the nuclear periphery, leads to the disruption of nuclear peripheral heterochromatin and clustering of chromatin at the nuclear pore (Eriksson *et al* 2003, Goldman *et al* 2004, McCord *et al* 2013). Interestingly, the loss of heterochromatin in progeria appears to contradict the observation of H3K9me3 and heterochromatin protein 1 (HP1) senescence-associated heterochromatic foci (SAHF) formation in senescent cells (Narita *et al* 2003). However, Hi-C indicates that the chromatin architecture is similar in ageing and progeria models: local connectivity was lost with an accompanying gain in inter-TAD interactions, indicating a loss of internal structure and heterochromaticity (Chandra *et al* 2015). Local interactions were markedly reduced in a step-wise manner when comparing between embryonic stem cells, somatic, and senescent cells (Chandra *et al* 2015), suggesting that senescence might be the end-product of a continuous remodelling process where TADs gradually lose integrity.

A large proportion of non-coding variants from genome-wide association studies (GWAS) overlap DHS, which are putative regulatory elements, but their

involvement in the disease of interest is unclear. Integrative analyses with high resolution contact mapping have shed some light on their relevance. By combining contact maps from cortical and germinal brain tissues, Won *et al* (2016) were able to link 108 schizophrenia risk variants with transcription factors involved in neurogenesis and cholinergic signalling pathways, and identified an enhancer that regulates *FOXG1*, which has previously been implicated as a risk gene. Similarly, this approach has been repeated in other systems to annotate and connect non-coding loci and single-nucleotide polymorphisms (SNPs) with genetic elements such as protein-coding genes and long non-coding RNA (lncRNAs) (Hughes *et al* 2014, Dryden *et al* 2014, Jäger *et al* 2015, Mifsud *et al* 2015, Martin *et al* 2015, Corradin *et al* 2016, Javierre *et al* 2016). Disease-associated SNPs are also often found in frequently interacting enhancer regions (FIREs) (Schmitt *et al* 2016). Crucially, these studies successfully impugn the assumption that disease risk genes and their corresponding variants have to be in close linkage disequilibrium (Mishra & Hawkins 2017). The various examples supplied in this section argue strongly for a role of genome organisation at various hierarchical levels in function (e.g. maintaining appropriate regulatory contacts) and disease.

Mechanisms of chromatin folding

A comprehensive understanding of the mechanisms underlying the construction of chromatin architecture is lacking. Contemporary models

promulgate loop extrusion as the means by which chromatin loops are formed (Alipour & Marko 2012, Sanborn *et al* 2015, Fudenberg *et al* 2016, Goloborodko *et al* 2016, Gassler *et al* 2017). Broadly, loops are formed through the interplay between loop extruding factors (LEFs) and boundary elements (BEs) that interfere with progressive loop extrusion. Each LEF such as cohesin binds chromatin at adjacent regions and extrudes loops by translocating along DNA; translocation is delayed or impeded by BEs, which underlies the formation of TADs (Sanborn *et al* 2015, Fudenberg *et al* 2016). BEs likely do not form stable loops as hypothesised in Rao *et al* (2014) and polymer models making this assumption fit poorly with Hi-C data (Fudenberg *et al* 2016). In accordance with the model, continuous loop extrusion within the confines of TADs result in frequent intra-TAD interactions (Fudenberg *et al* 2016, Andrey & Mundlos 2017) and significant interactions within TADs could indicate more permeable BEs vis-à-vis TAD boundaries.

CTCF binding sites are extensively studied candidates for BEs. CTCF is enriched at TAD boundaries (Dixon *et al* 2012, Yaffe & Tanay 2011) and TAD corner-peak foci (Rao *et al* 2014), has a relatively high residency on chromatin (Nakahashi *et al* 2013), and CTCF depletion reduces insulation at TAD boundaries (Zuin *et al* 2014, Nora *et al* 2017). The CTCF binding motif is non-palindromic and can thus be assigned a direction (Kim *et al* 2007). Importantly, chromatin loops are preferentially formed between convergent CTCF sites (Rao *et al* 2014, de Wit *et al*

2015, Nichols *et al* 2015, Rudan *et al* 2015), hinting that CTCFs could function as directional BEs. This is supported by genome editing studies, where the inversion of a CTCF binding site caused neighbouring TADs to merge (de Wit *et al* 2015, Guo *et al* 2015). Besides CTCFs, BEs can theoretically be any feature that interferes with LEF translocation, such as a high occupancy of chromatin proteins (van Bortle *et al* 2014) or very active genes bound by bulky transcriptional machinery (Ulianov *et al* 2016) which are enriched at TAD borders (Dixon *et al* 2012).

Cohesin is a prime candidate for LEFs. It binds to the C-terminus of chromatin-bound CTCF that faces in toward TADs (Xiao *et al* 2011). Moreover, it is known to translocate along chromatin (Stigler *et al* 2016), is structurally similar to known motor proteins (Nasmyth *et al* 2000), and is thought to be able to form chromatin loops (Alipour & Marko 2012). Indeed, condensin, a structural maintenance of chromosomes (SMC) complex similar to cohesin, was shown to extrude loops processively via real-time microscopy in *Saccharomyces cerevisiae* (Ganji *et al* 2018). That said, the estimated rate *in vivo* at which loops are extruded (~375 bp/s) (Rao *et al* 2017) is not congruous with *in vitro* studies of cohesin translocation, where DNA sliding is estimated at 1-2 bp/s (Stigler *et al* 2016).

Transcription is another candidate for loop extrusion; RNA polymerase II tracks along DNA with cohesin (Jonkers & Lis 2015, Davidson *et al* 2016). Indeed,

while select enhancer-promoter interactions and TADs remain relatively stable after transcription inhibition (Palstra *et al* 2008, Hug *et al* 2017, Ke *et al* 2017), they may augment the speed of extrusion (Dekker & Mirny 2016), as evinced in auxin-degron experiments where the recovery of cohesin-dependent loop domains is faster when they span highly active super-enhancers (Rao *et al* 2017). Transcription-induced supercoiling may also help push cohesin rings along (Racko *et al* 2018). Moreover, there is a recent proposal suggesting the possibility of osmotic pressures and 1D diffusion driving cohesin translocation (Brackley *et al* 2018).

A growing body of evidence indicates that loop extrusion and compartmentalisation are antagonistic processes (Nuebler *et al* 2018). The loss of cohesins led to weaker TADs but strengthened compartmentalisation (Schwarzer *et al* 2017, Rao *et al* 2017, Wutz *et al* 2017, Haarhuis *et al* 2017), creating sharper transitions (Rao *et al* 2017) and finer compartments which align better with epigenetic markers than wild-type compartments (Schwarzer *et al* 2017). In contrast, increasing cohesin residency through the depletion of the cohesin unloader WAPL resulted in longer and more chromatin loops, stronger TADs, and weaker compartments (Haarhuis *et al* 2017, Wutz *et al* 2017). Lastly, the removal of BEs such as CTCF did not impact compartmentalisation or loop extrusion significantly (Nuebler *et al* 2018) but led to a loss of TAD boundary insulation

(Nora *et al* 2017). In all, contact maps reflect the competition between at least two modes of chromatin folding: loop extrusion and a separate mechanism that is likely due to the propensity for euchromatic or heterochromatic proteins to self-associate, such as the polyhomeotic proteins in PRC1 that are capable of self-interaction (Isono *et al* 2013), with phase separation driving compartmentalisation (Larson *et al* 2017, Strom *et al* 2017, Nuebler *et al* 2018).

Methods to interrogate regulatory interactions

Each technique suffers from limitations depending on the research question. Hi-C is adept at capturing large-scale structures but is less efficient for finer architecture such as EPIs. Early experiments found TADs and compartments with as few as 8.4 million read pairs (Lieberman-Aiden *et al* 2009), but only around 10,000 chromatin loops were called with around 5 billion contacts (Rao *et al* 2014), although this is contingent on the method for calling significant interactions (Forcato *et al* 2017). Hi-C maps are typically low resolution. To increase the resolution of Hi-C libraries linearly, sequencing depth has to be increased quadratically, which can make Hi-C studies cost-prohibitive (Schmitt *et al* 2016).

Strategies to circumvent this shortcoming typically involve subsetting the genome. Capture-C (Hughes *et al* 2014, Davies *et al* 2016), Capture Hi-C (Schoenfelder *et al* 2015a, Schoenfelder *et al* 2018), Targeted Chromatin Capture

(T2C) (Kolovos *et al* 2014), and targeted DNase Hi-C (Ma *et al* 2018) rely on probes to enrich for coverage at regions of interest. Alternative approaches adopted by chromatin interaction analysis with paired-end tag (ChIA-PET) (Fullwood *et al* 2009, Li *et al* 2017), proximity ligation-assisted ChIP-seq (PLAC-seq) (Fang *et al* 2016), or HiChIP (Mumbach *et al* 2016) interrogate interactions at proteins of interest such as histone marks (Heidari *et al* 2014, Fang *et al* 2016), CTCF and RNA pol II (Tang *et al* 2015), YY1 (Weintraub *et al* 2017). However, these approaches survey a predetermined interaction space, which require *a posteriori* knowledge of the genetic elements or proteins used. In addition, ChIA-PET and HiChIP need ChIP-grade antibodies.

C. elegans genome and chromatin

C. elegans worms are predominantly hermaphrodites and males arise by spontaneous non-disjunction in the germline (~0.1%) (Altun & Hall 2009); hermaphrodites self-fertilise for homozygous worms to produce genetically identical progenies. Its life cycle comprises an embryonic stage, four larval stages (L1-4) corresponding to molts at the end of each stage, and adulthood.

C. elegans shares many chromatin features with other eukaryotes. Histone modifications are similar to that in other animals (Kolasinska-Zwierz *et al* 2009,

Liu *et al* 2011, Ho *et al* 2014) and the worm has many homologs of human chromatin proteins (reviewed in Cui & Han 2007). Interestingly, it lacks DNA methylation and known insulator proteins (Ong *et al* 2009), in particular CTCF (Heger *et al* 2009) like in yeast and plants, which raises the question of how chromatin is organised in the worm.

The worm genome is made up of five autosomes (I to V) and a sex chromosome (X). The chromosomes are holocentric and do not have clearly defined centromeric regions but the worm still shares a similar kinetochore machinery and cell division mechanisms with other eukaryotes (Kitagawa 2009). However, autosomes can still be segmented into a central region and flanking arms based on transitions in meiotic recombination rates (Barnes *et al* 1995), with the arms having higher rates. The arms are also enriched for tandem and inverted repeats, fewer essential genes, and lower gene activity (Kamath *et al* 2003, Prachumwat *et al* 2004), likely as result of being tethered to the nuclear periphery (Ikegami *et al* 2010).

Regulatory elements are not well mapped and annotated in the worm. This is because messenger RNA (mRNA) for around 70% of *C. elegans* genes are trans-spliced, wherein the 5'-ends of pre-mRNA are trimmed and replaced with a 22-nucleotide spliced leader (SL1 or SL2) that is contributed by 100-nucleotide small

nuclear ribonucleoprotein particles (reviewed in Hastings 2005). Degradation at the 5'-end obscures the transcription start site if one relies of analyses of mature mRNA. To circumvent this issue, studies have relied on assaying transcription initiation from nuclear RNA or by inhibiting trans-splicing (Chen *et al* 2013, Kruesi *et al* 2013, Saito *et al* 2013). Alternatively, regulatory elements have been mapped with DNase-seq and ATAC-seq data, and annotated based on proximity to exons or chromatin states (Daugherty *et al* 2017, Ho *et al* 2017).

Active regulatory elements - promoters and enhancers - are known to have different chromatin states. Transcription initiation occurs at both promoters and enhancers, and many of them experience divergent, bidirectional initiation from two independent sites (Koch *et al* 2010, Chen *et al* 2013). Promoters typically have high H3K4me3 and low H3K4me1 levels, whilst the opposite is observed for enhancers (Heintzman *et al* 2009). However, later studies in human and flies indicated that H3K4me1/3 levels correlate with levels of transcription activity at the corresponding element, rather than their identity as a promoter or enhancer (Core *et al* 2014). In addition, developmental genes can lack H3K4me3 despite being actively transcribed (Zhang *et al* 2014).

Practically, promoters and enhancers can be functionally distinguished by their transcriptional outputs. Whilst transcription occurs for both types of

elements, promoters produce stable transcripts that are subsequently processed, but enhancers typically produce short, unstable transcripts (Core *et al* 2014). Based on this principle, 15,714 protein-coding promoters and 19,231 putative enhancers were annotated based on ATAC-seq and RNA-seq data in several developmental stages by other members of the laboratory (Jänes *et al* 2018). Promoters were defined as elements having significant transcription elongation originating from the element in at least one direction and one developmental stage (Jänes *et al* 2018). Enhancers were defined by the presence of transcription initiation but the absence of an elongation signal (Jänes *et al* 2018). The mapping on these elements in Jänes *et al* 2018 now allows me to address the question of how regulatory elements communicate with each other in 3D to control gene expression.

The first Hi-C map in *C. elegans* reported no canonical TADs on the autosomes but observed insulated chromatin interaction domains on the X chromosomes (Crane *et al* 2015). Domain boundaries correspond to dosage compensation complex (DCC) recruitment sites. In hermaphrodites, DDC, a condensin-I-like complex, organises chromatin architecture in the X chromosomes, resulting in the down-regulation of X chromosome genes by approximately half for each chromosome (Meyer 2010). DCC is recruited onto the X chromosomes by SDC-2 (Albritton *et al* 2017) at recruitment element on the X (*rex*) sites and significantly strong interactions occur between pairs of the top 25

rex sites (Crane *et al* 2015). These domains and interactions between *rex* sites are dependent on DCC and are lost in DCC-defective worms (Crane *et al* 2015).

Much less is known about how autosomes are organised. Given the coarse resolution of this contact map (10-50 kb) and the compactness of the *C. elegans* genome with more than 20,000 protein-coding genes in 100 Mb of genomic space, it is unclear if smaller domains similar to globules in *S. pombe* exist (Mizuguchi *et al* 2014).

Project objectives

Work in cell culture has been instrumental in shaping our current knowledge of nuclear architecture and will continue to provide new insights. However, *in vivo* work in animals is crucial for a fuller understanding and particularly for determining how chromatin is regulated in development. *C. elegans* provides an outstanding developmental system for addressing this due to its rapid development (wild-type worms reaches adulthood in about 3 days), ease of culture, small well-annotated genome (~100 Mb), and conserved chromatin features and regulators.

To date, genome organisation in the worm has not been well-studied. Beyond the DCC-mediated domains on the X chromosome, little is known about how the autosomes are organised. It is still an open question as to whether the worm

autosomes have contact domains and compartments. An earlier study (Crane *et al* 2016) could have missed them due to their use of low resolution assays (10-50kb) and the compact nature of the worm genome (i.e. about 20,000 genes in 100Mb of genomic space). Moreover, whilst regulatory elements have recently been mapped and annotated (Jänes *et al* 2018), enhancers are still assigned to their putative target genes based on proximity in linear distance and not physical distance. Because interactions between regulatory elements are extensive and relevant for controlling gene expression in other systems, it is important to understand if they exist in the worm. If so, their identification is necessary for understanding gene regulation.

Here, I set out to map the 3D interactome in *C. elegans*. Particularly, to ask if long-range EPIs exist in the worm, and if so, the extent to which they participate in the regulation of gene expression (as in mammalian systems). Beyond EPIs, I want to know if higher order structures such as interaction hubs/networks, TADs and compartments exist in the worm, and their relationship with gene expression. In addition, I endeavour to identify potential protein candidates that could be implicated in regulating 3D architecture.

To tackle these issues, I set out to develop a novel method based on Hi-C that could comprehensively study genome organisation in the worm at multiple

scales. Regulatory element interactions have primarily been mapped using methods targeting particular elements (e.g. promoters) because they are generally not visible in whole genome Hi-C maps. To enable global mapping of such interactions at high resolution, I developed the Accessible Region Conformation-Capture (ARC-C). ARC-C enriches for interactions between regulatory elements, which allows me to define the landscape of significant chromatin interactions between regulatory elements at 500 bp resolution. However, ARC-C also provides information at non-regulatory regions, which allows for studies of domains and compartments at lower resolution.

Moreover, with the enrichment at regulatory elements, I could screen for homotypic proteins that are enriched at both interaction ends, which suggests a putative function for these proteins in the mediation of these significant interactions. Thereafter, with the list of candidates, I applied ARC-C to investigate the role of a significantly enriched transcription factor (BLMP-1) in chromatin looping. At the domain and compartment level, I used ARC-C to question the role of H3K9 methylation in contact domain and compartment formation.

An important caveat to note here is that the ARC-C, modENCODE, and other in-house data used in this thesis came from whole animals. Observations and conclusions are thus done on a population level and may not reflect events in

single cells. In particular, since whole animals were used, representation from different tissues will not be equal and signal from smaller tissues may be diluted in aggregate studies. Interpretations for tissue-specific genes are valid, but will have to be verified in sorted or purified samples for the corresponding tissues - this is an area that is being worked on by others in the lab.

In this thesis, I report, for the first time, the existence of insulated contact domains and compartments on the autosomes in *C. elegans*. Importantly, the formation or maintenance of regulated compartments (i.e. B compartments) in the worm relies on H3K9 methylation. With the sensitivity afforded by ARC-C, I was able to generate a list of putative factors that could be involved in chromatin looping, which is important because only a few such factors are known so far.

Chapter I describes the development of ARC-C and the means to process and analyse ARC-C data. **Chapter II** characterises significant interactions in wild-type worms. **Chapter III** identifies proteins that are enriched at interaction ends and includes results from *blmp-1* mutants. **Chapter IV** describes domains and compartments and the role of H3K9 methylation in the mediation of these structures. **Chapter V** details steps undertaken toward an improved version of ARC-C.

Scientific acknowledgements

I performed all the experimental work except for the generation of three hub deletion strains, which were made by Chiara Cerrato and Andrea Tufekcic, and hub deletion experiments, which were done with help from Garima Sharma and Diljeet Gill. Ni Huang performed most of the bioinformatics analyses and we collaborated on the design of analyses. I characterised regulatory interactions (i.e. promoter-promoter, promoter-enhancer) and hubs, did the analyses on condensin, cohesin, and other chromatin regulators.

CHAPTER I: ARC-C DEVELOPMENT AND EVALUATION

Accessible region conformation-capture (ARC-C) aims to interrogate mulitple levels of 3D genome organisation. To study regulatory interactions, which is relatively challenging, ARC-C enriches for interactions at regions of open chromatin that reflects DNase I hypersensitivity. In this chapter, I outline the conceptual bases of ARC-C and discuss the way we process, normalise, and call significant interactions from ARC-C data. Thereafter, I compare and evaluate ARC-C with a published Hi-C in wild-type *C. elegans* late embryos (Crane *et al* 2015).

Principles underlying ARC-C

Hi-C has two major limitations when used to study regulatory interactions, namely a theoretical restriction-fragment resolution and highly complex libraries that make high resolution analyses cost-prohibitive. The availability and position of restriction cut sites affect the ability to analyse particular loci. Restriction cut site motifs are depleted near regulatory elements (as defined in Jänes *et al* 2018; the mapping and annotation of these elements will be discussed at length later in

this section) (**Fig 1.1**). As a result, genetic elements are often conflated within individual restriction fragments, which could lead to an inability to interrogate certain regulatory regions of interest at high resolution.

Of all possible *DpnII* restriction fragments with DNase I hypersensitive sites (DHS), around 17.7% in wild-type L3 worms, 11.0% in *Drosophila* S2 cells, and 15.9% in human K562 cells have more than one DNase I hypersensitive site (**Fig 1.2**), potentially making gene assignment ambiguous. The sensitivity of restriction-enzyme based C-methods at regulatory elements is further undermined by the number of restriction fragments that do not contain any regulatory elements (87.2% of *DpnII* fragments in *C. elegans*, 98.4% in *Drosophila* S2, and 97.0% in human K562). Therefore, some form of enrichment would be required to better study regulatory interactions.

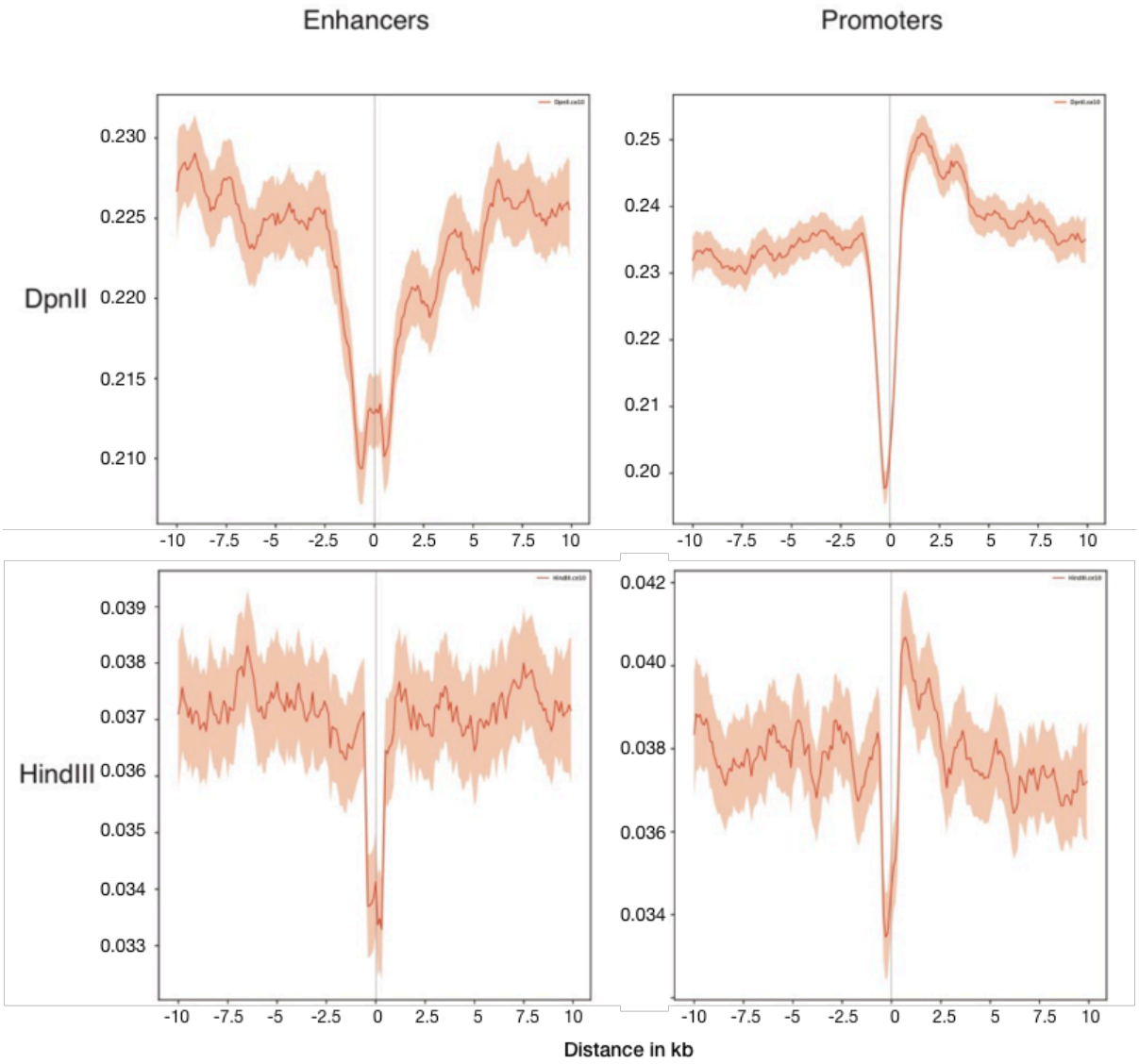
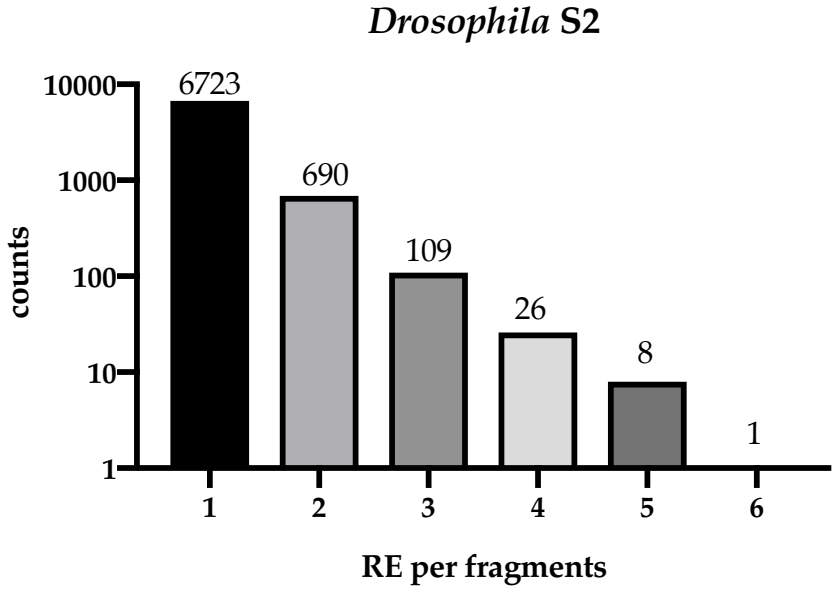
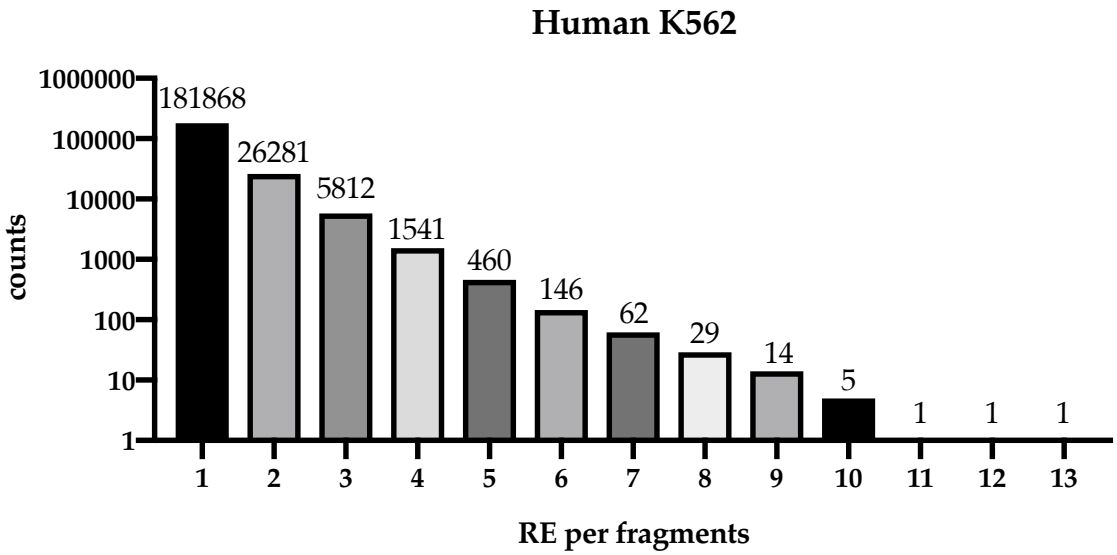


Figure 1.1: Aggregate distribution of restriction enzyme cut sites centred over regulatory elements in *C. elegans.*


Drosophila S2
10000
1000
100
10
1
counts
6723
690
109
26
8
1
1
2
3
4
5
6
RE per fragments


Human K562
1000000
100000
10000
1000
100
10
1
counts
181868
26281
5812
1541
460
146
62
29
14
5
1
1
1
1
2
3
4
5
6
7
8
9
10
11
12
13
RE per fragments

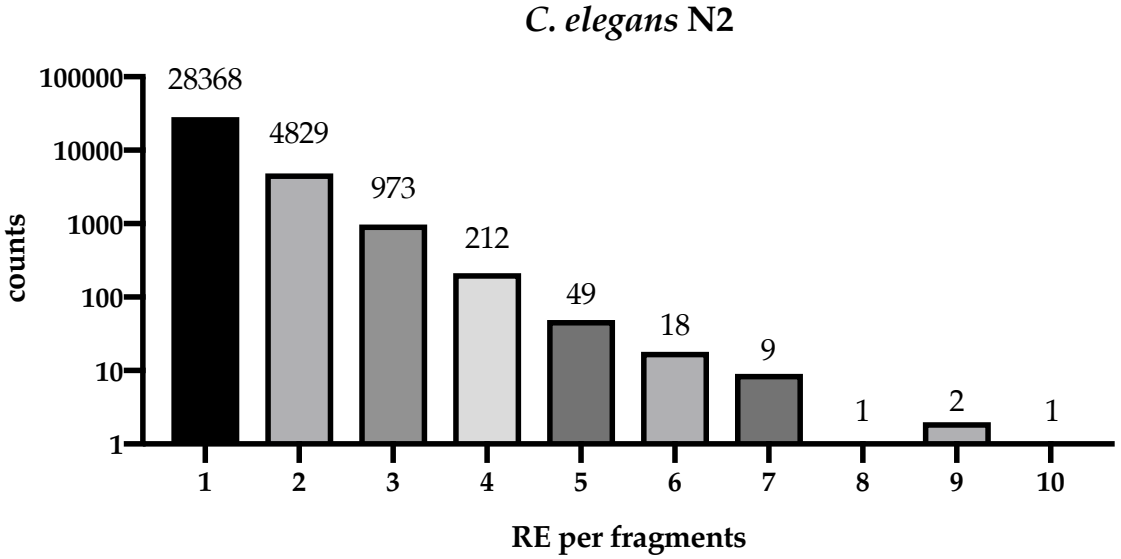


Figure 1.2: Count of *DpnII* restriction fragments with varying number of DHS in *Drosophila* S2, human K562, and *C. elegans* wild-type (N2). Genomes were computationally digested with *DpnII* to obtain fragments and DHS from DNase-seq and ATAC-seq were assigned to each fragment.

ARC-C adopts the principles of *in-situ* Hi-C (Rao *et al* 2014), DNase I hypersensitivity site mapping (Boyle *et al* 2008), and transposase-mediated profiling of accessible chromatin (Buenrostro *et al* 2013) (**Fig 1.3**) to enrich for interactions between regulatory elements in the genome (**Method**). *In-situ* or in-nucleus proximity ligation has the benefits of reducing noise from spurious ligations, improving reproducibility and reducing biases (Gavrilov *et al* 2013, Rao *et al* 2014, Nagano *et al* 2015).

Instead of using restriction enzymes to digest the genome, ARC-C uses DNase I *in situ*. In the same vein as a previously published DNase-Hi-C (Ma *et al*

2015, Deng *et al* 2015, Ramani *et al* 2016), the use of DNase I allows us to overcome the theoretical resolution limit imposed by restriction enzymes. Unlike the existing DNase-Hi-C, which digests the genome uniformly, I apply DNase I at a concentration that preferentially, but not exclusively, digests the genome at nucleosome-depleted regions, similar to DNase-seq (Song & Crawford 2010), thus enriching for cuts at regions of open chromatin. These regions are typically active *cis*-regulatory elements such as promoters, enhancers, insulators, and silencers (Thurman *et al* 2012). Subsequently, informative interactions are captured by an *in situ* tagmentation reaction mediated by a hyperactive Tn5 transposase, which preferentially targets open chromatin (Buenrostro *et al* 2013).

ARC-C

Experimental steps

Chromatin is first fixed in place with formaldehyde (**Fig 1.3**). DNase I is applied to lightly digest the chromatin *in situ*. Practically, I make three separate ARC-C libraries from different concentrations to account for slight day-to-day variations in DNase I digestion; the optimal library is selected for sequencing after several quality control steps, which include a qualitative evaluation of the extent of DNase I digestion with electrophoresis and quantitative polymerase chain reaction (qPCR) measurement of a diagnostic DHS (discussed later in this chapter). Overhangs from DNase I digestion were blunted and proximity ligation

was performed subsequently. Informative interactions that are represented by chimeric DNA fragments were enriched for and captured with Tn5 transposase, which fragments DNA and attaches Illumina sequencing adaptors at both ends of the fragments. Libraries were then strictly size-selected for a maximum insert size of 600 bp with solid phase reverse immobilisation beads. In all, ARC-C can be completed in 2 days.

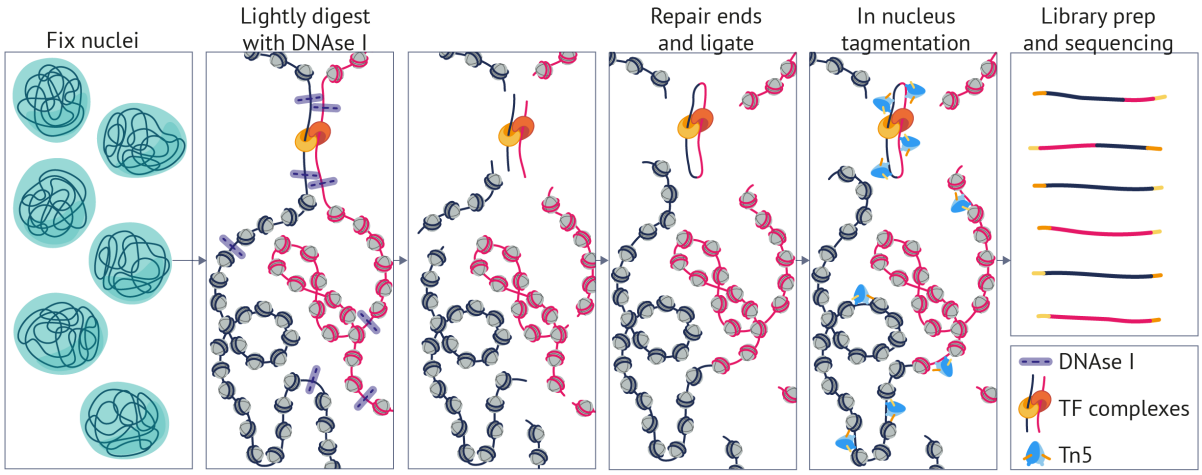


Figure 1.3: Schematic of ARC-C protocol.

Data processing

We map reads individually, filter raw data based on mapping quality (MAPQ ≥ 30), remove PCR artefacts, mitochondrial DNA (mtDNA), and blacklisted regions, then pair the reads - these constitute 'valid reads' (**Fig 1.4**). The chosen size cutoff represents the theoretical resolution of our technique and allows us to filter for 'informative reads': in principle, read pairs that map at a distance greater than the size cutoff (600 bp) are likely to have undergone DNase I

digestion and proximity ligation (**Fig 1.4**). Concomitantly, we look at the orientation of paired reads (forward-forward, forward-reverse, reverse-reverse, reverse-forward) to corroborate our size cutoff. Without digestion and proximity ligation, read pairs are predominantly in a forward-reverse orientation. After digestion, DNA fragments have a theoretically equal probability of re-ligating in all four permutations of orientation. The fragment size at which this occurs can help inform the size cutoff threshold but is typically consistent with the physical size-selection cutoff. **Fig. 1.5** shows an example from a wild-type L3 larval stage ARC-C library: "**N2_1**". In agreement with a physical size cutoff of 600 bp, the proportion of forward-reverse read pairs sharply drops off to approximately the same levels as the other orientations (**Fig 1.5**). We use these informative reads to call significant interactions or to construct contact maps (**Fig 1.4**).

Unlike classical Hi-C, there is typically a steep drop from valid to informative reads (e.g. 207 to 15 million; **Fig 1.4**) for ARC-C. This is primarily because ARC-C, in its current form, lacks a step to enrich for informative junctions, which is done in Hi-C through a pull-down of biotinylated nucleotides using streptavidin beads. This shortcoming is addressed in later iterations of ARC-C, as discussed in **Chapter V**.

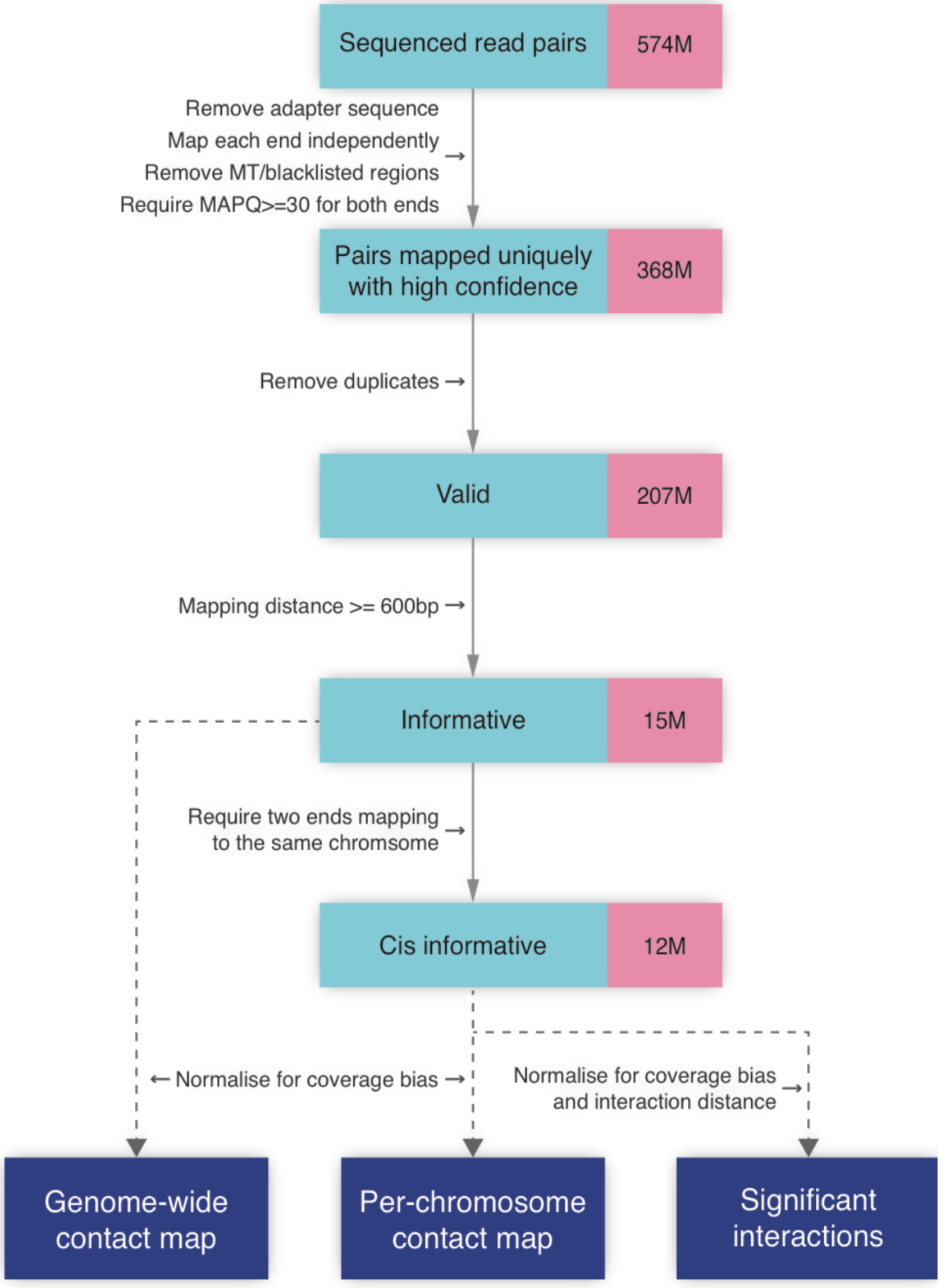


Figure 1.4: Schematic of ARC-C data processing steps. Red boxes indicate the number of read pairs at each step of the procedure.

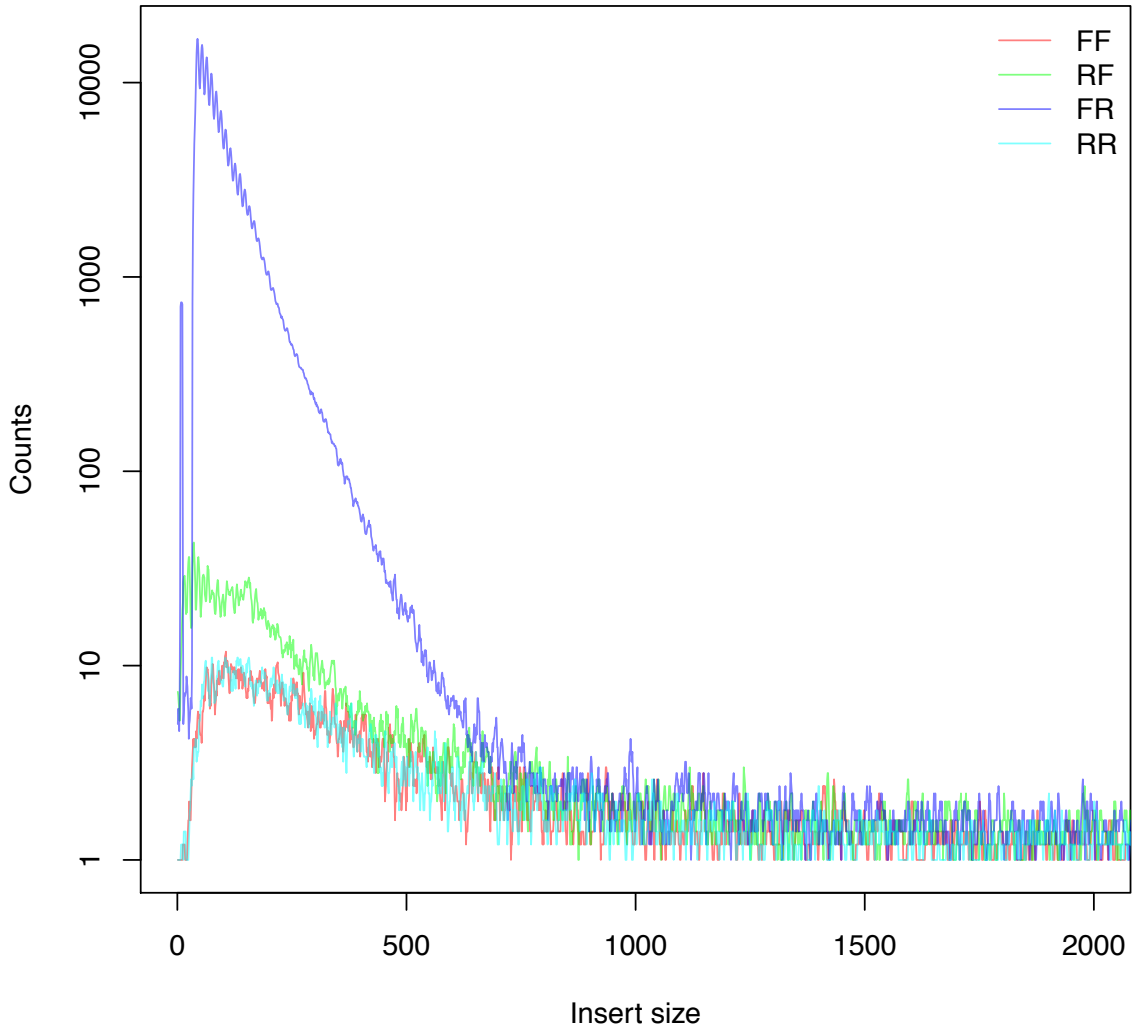


Figure 1.5: Counts from representative library "N2_1" of inserts of each mapping orientation - forward-forward (FF), reverse-forward (RF), forward-reverse (FR), reverse-reverse (RR) - at varying insert sizes. FR represents the natural orientation of read pairs from undigested and religated sequencing libraries.

Evaluating ARC-C data

ARC-C libraries were made from wild-type Bristol N2 *C. elegans* L3 stage larvae. I sequenced 4 libraries comprising 3 biological and 2 technical replicates (**Table 1.6, Appendix - Table A1.1**). Biological replicates - N2_1*, N2_2*, N2_3* - were libraries made from separate collections of worm larvae (**Table 1.6**).

Technical replicates were libraries made on separate days (N2_2a/b) from the same worm collection (**Table 1.6**). Informative interactions were binned at different intervals - 1 kb, 2 kb, 5 kb, 10kb, 20kb, 50kb - and tested for correlation. Pearson correlation can be overly sensitive to outliers (Pernet *et al* 2013) and its performance is understandably dependent on resolution (Yardimci *et al* 2018). The complexity of C-based libraries and data sparcity at high resolution (bins having zero or few reads) can create the impression of outliers, reducing the overall correlation coefficient. I see a similar effect with a Pearson correlation of 0.96-0.99 at 20kb resolution and 0.61-0.88 at 1kb resolution (**Fig 1.7**). However, these values can be considered well-correlated (e.g. biological replicates had a Pearson correlation of 0.985 at 50kb resolution in Crane *et al* 2015). I therefore pooled replicates for N2 L3 larvae for statistical rigour. In all, I obtained about 414 million valid and 24.5 million informative reads.

Surprisingly, the correlation coefficient can be lower between technical replicates (N2_2a and N2_2b) than biological replicates (e.g. N2_1 and N2_2a). This can plausibly be attributed to the intrinsic variability and sensitivity of DNase I digestion: despite using the same biological material, slight day-to-day variation in temperature and time can result in chromatin being cut differently. This difference may also be exaggerated by the use of Pearson correlation as a measure of reproducibility.

Libraries	Enrichment over DHS	Valid reads	*Cis* Informative reads	*Cis* ratio
N2_1	3.7	109,921,232	7,253,030	77.81%
N2_2a	3.9	89,174,244	6,034,950	78.07%
N2_2b	3.4	134,038,220	6,336,580	77.87%
N2_3	5.4	80,576,176	4,760,748	78.52%

Table 1.6: Key statistics for wild-type L3 libraries. Enrichment over DHS measures the enrichment of informative read coverage at DHS over background, which reflects signal to noise. Cis (%) measures the percentage of *cis* informative reads over all informative reads.

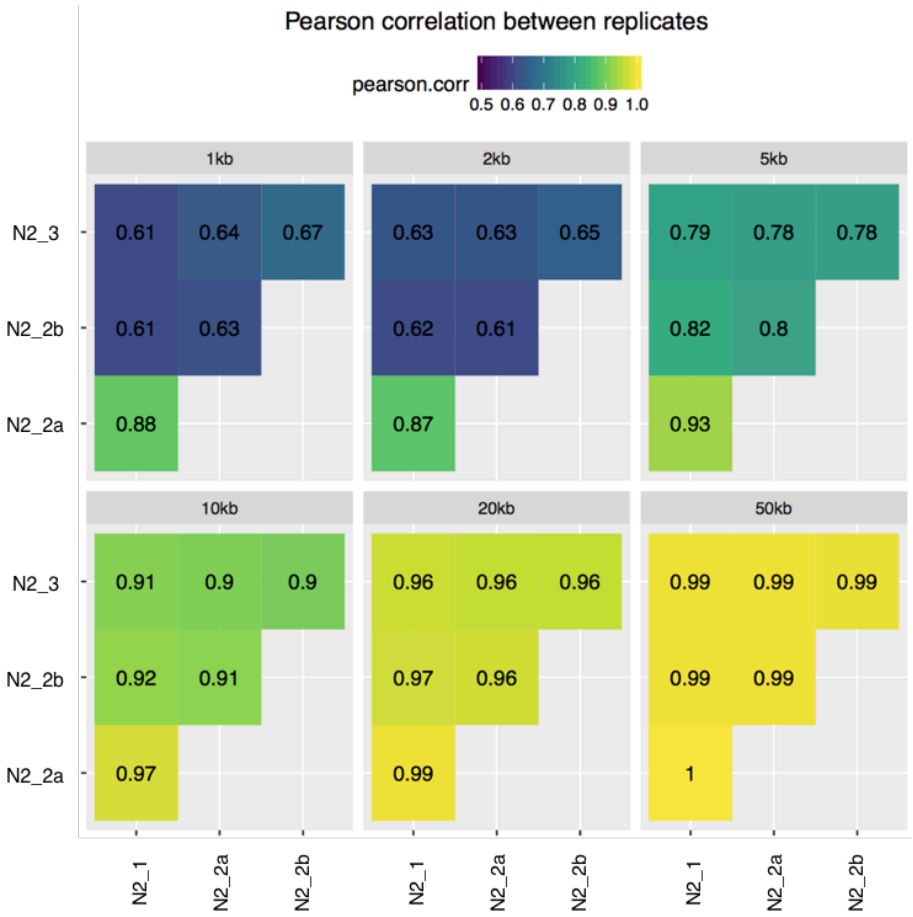


Figure 1.7: Pearson correlation between wild-type L3 ARC-C library replicates binned at 1kb , 2kb, 5kb, 10kb, 20kb, and 50kb resolution.

Conceptually, ARC-C peaks correspond to 'baits' - significant interactions are typically between peaks. I therefore compared the coverage of ARC-C valid reads and ATAC-seq normalised reads in the same strain and larval stage to ensure that there are no new and unexpected 'baits' in ARC-C which would result in false positives. I also compared a *C. elegans* mixed embryos Hi-C dataset from Crane *et al* to validate ARC-C's ability to enrich for informative reads over open chromatin. Hi-C aims to achieve equal representation of the genome by digesting chromatin uniformly and thoroughly. This manifests as an even coverage of informative read pairs (Imakaev *et al* 2012). As expected, the coverage for informative reads in Crane Hi-C was fairly even (**Fig 1.8**), but the coverage for ARC-C and ATAC-seq form enriched peaks over DHS (**Fig 1.9**). Peak positions were roughly concordant and of similar strength (**Fig 1.8**). I show here that ARC-C libraries from N2 L3 larvae are reproducible and have similar 2D coverage with ATAC-seq libraries, indicating that the method is robust and functional.

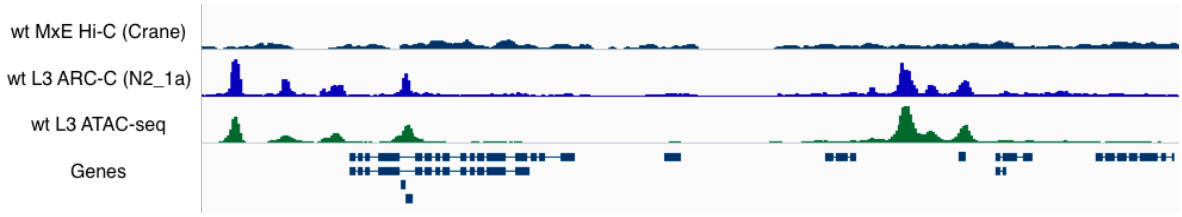


Figure 1.8: Top to bottom - coverage of HiCUP-filtered informative Hi-C reads from wild-type mixed embryos (Crane *et al* 2015), wild-type L3 stage ARC-C valid reads, wild-type L3 stage ATAC-seq normalised reads, and genes. chr I: 6,075,000 - 6,100,000.

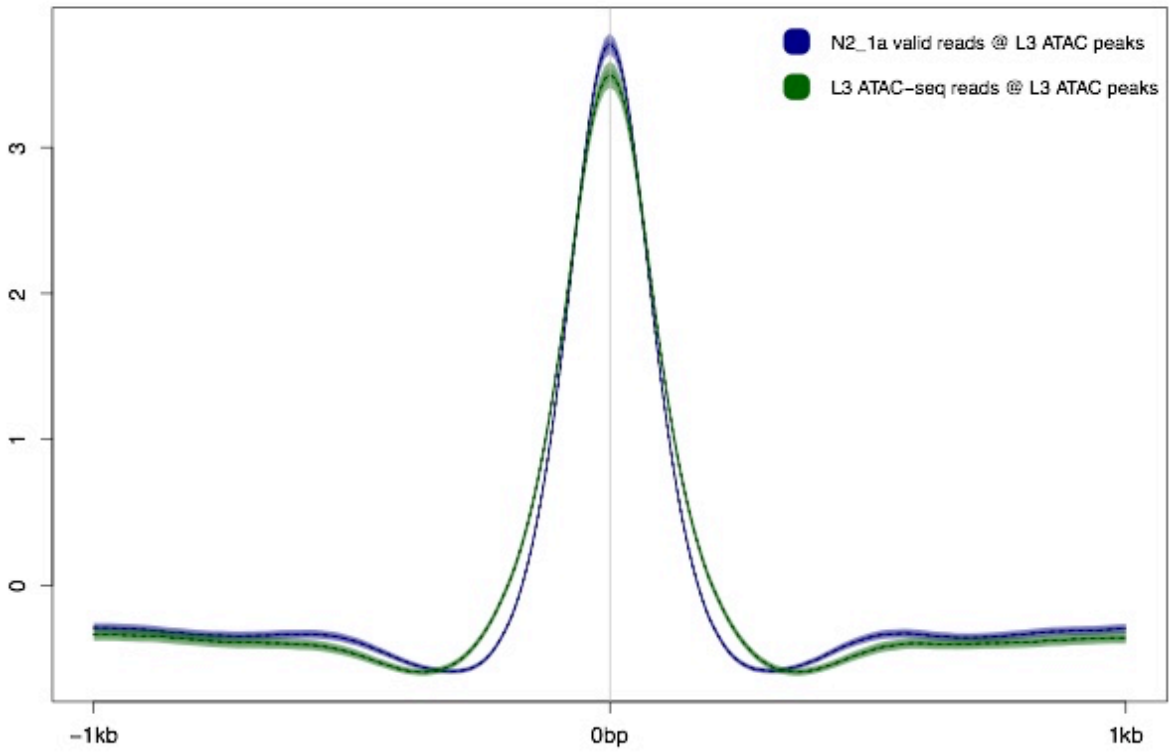


Figure 1.9: Aggregate coverage of N2_1a ARC-C valid reads (blue) and L3 ATAC-seq normalised reads (green) centred on L3 ATAC-seq peaks.

I used a set of diagnostic metrics to assess the quality of ARC-C libraries - the extent of DNase I digestion, the aggregate coverage of informative reads over DHS, the peak height of a diagnostic DHS as quantified by qPCR, and *cis* percentage of informative reads. Quality control metrics that are pertinent to sequencing libraries in general - for example, mitochondrial DNA content, adaptor content or mapping quality - but not unique to ARC-C libraries will not be discussed here.

The signal at DHS is highly sensitive to the extent of DNase I digestion, similar to DNase-seq (Song & Crawford 2010), with heavier digestion typically

resulting in lower signal to noise. The optimal level of DNase I digestion has to be determined empirically for different cell types. I found that the DNase digestion pattern at 50U/ml and 100U/ml were typically optimal for L3 stage *C. elegans* larvae (**Fig 1.10**). I made L3 stage ARC-C libraries for a series of DNase I concentrations (5, 10, 25, 50, 100, 200U/ml). 50U/ml and 100U/ml corresponded to the highest coverage enrichment over DHS (**Table 1.11**). Coverage enrichment over DHS measures the enrichment of aggregate single-ended informative read coverage over background at wild-type L3 ATAC-seq peaks, which reflects the genome-wide signal to noise. The qualitative evalution of the extent of DNase digestion was applied to all L3 stage ARC-C libraries and I typically make three libraries concurrently (usually 25U/ml, 50U/ml, 100U/ml) to account for slight variations in DNase digestion that could substantially alter the digestion patterns.

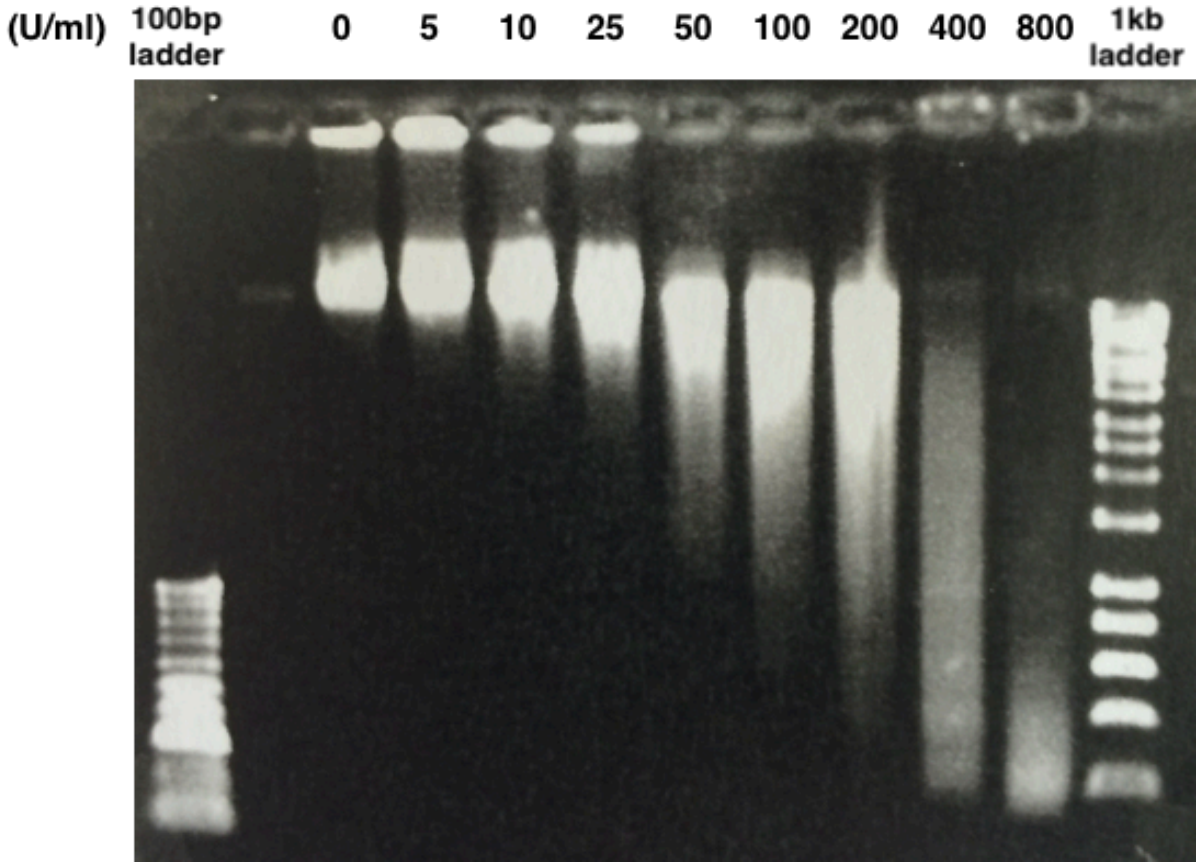


Figure 1.10: Gel electrophoresis of DNase I digested chromatin on 1% agarose gel. Units (U/ml) correspond to final concentrations of DNase I used.

Extent of DNase I digestion (U/ml)	Coverage enrichment over DHS
5	1.3
10	1.5
25	2.2
50	3
100	3.7
200	2.8

Table 1.11: Aggregate coverage enrichment of informative reads over DHS genome-wide of ARC-C libraries at different extents of DNase I digestion.

Good ARC-C libraries have high enrichment of informative read coverage over DHS. Before paired-end sequencing, I conducted an additional quality control step and tested the enrichment of a diagnostic DHS over background via qPCR. I selected the promoter of the essential gene *gap-3*, which was accessible at most *C. elegans* developmental stages (mixed embryos, L1, L2, L3, L4, young adults) (**Fig 1.12**). Empirically, ARC-C libraries that had a fold-change above 4 via qPCR typically turned out to be libraries with good signal to noise.

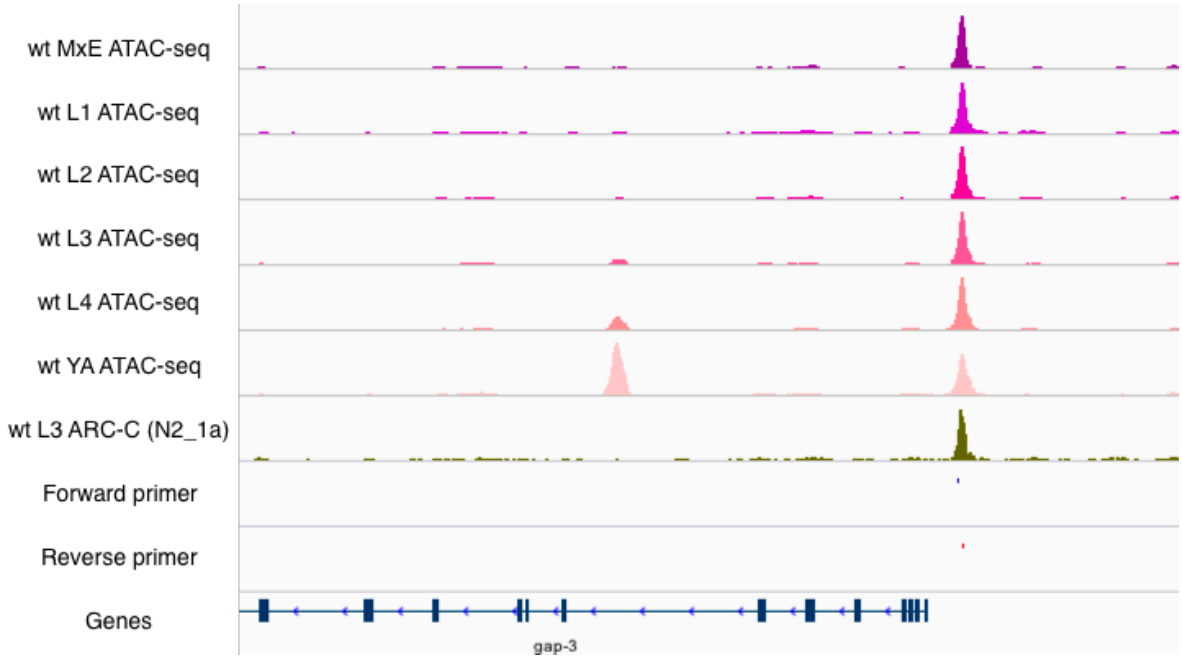


Figure 1.12: Top-bottom: wild-type ATAC-seq coverage at different developmental stages. ARC-C coverage at L3 stage. Forward and reverse qPCR primers for diagnostic DHS. Chr I: 2,000,000 - 2,015,000.

A majority of informative read pairs (60-80%) map to the reference genome in *cis* (**Table 1.6**), which means they map to the same chromosome. *Trans* read pairs that map to different chromosomes are conversely in the minority. This observation is consistent with the phenomenon of chromosome territories and with good Hi-C libraries (Nagano *et al* 2015). A low *cis-to-trans* ratio would indicate compromised nuclear integrity, which is reminiscent of pioneering Hi-C experiments where DNA was extracted and purified from nuclei and the ligation of fragments was done in extremely diluted conditions (Lieberman-Aiden *et al* 2009). This would result in a reduced sensitivity to capture loops, domain, and compartment structures. That said, whilst a low *cis-to-trans* ratio is undesirable for *in-situ* experiments because *trans* interactions are typically uninformative, the remaining *cis* read pairs could still be useful.

Comparison and evaluation with published Hi-C

A previously published Hi-C map in *C. elegans* mixed embryos (hereinafter named the 'Crane Hi-C') (Crane *et al* 2015) found chromatin interaction domains (~1Mb) on X chromosomes that resemble mammalian TADs. Domain boundaries correspond to high affinity condensin-I-like dosage compensation complex (DCC) recruitment elements on the X (*rex)* sites. These boundaries are also enriched with DPY-27, a condensin subunit unique to dosage compensation (condensin DC), and other condensin subunits and condensin-interacting proteins such as DPY-26,

SDC-2, SDC-3 and DPY-30 (Albritton *et al* 2017). In DCC mutants, domain boundaries on X chromosome become weaker or lost, implying a putative role for the condensin complex in organising the chromosome. On the autosomes, large multi-Mbs domains are observed, which corresponded to the segmentation of autosomes into flanking chromosome arms and a central region, as defined by early recombination studies (Barnes *et al* 1995). The chromosome arms and central region are functionally distinct, with the arms having lower gene density, lower average gene expression, an enrichment in repeats, and LEM-2 binding: a lamina-associated protein that is required for association with the nuclear periphery (Ikegami *et al* 2010).

To test if we could recapitulate the broad, biological features seen in Crane Hi-C, we binned ARC-C data into large 50kb windows, and compared the contact maps and insulation profiles. A genome-wide contact map for ARC-C is presented in **Figure 1.13**. Segmentation by chromosomal arms and the central region can be observed more strongly in chr I, II, and III, and weakly in chr IV and V. In addition, the central regions for each autosomes appear to intermingle to an extent based on the inter-chromosomal matrices, with chr I, II, and III having a stronger effect. At 50kb resolution, informative interactions in ARC-C shared a Pearson correlation of 0.92 with informative interactions in Crane Hi-C (**Fig 1.14**), implying that broad structures are similar. Using chr I as an example, arm-centre

domain transitions (grey arrows) were aligned between ARC-C and Crane Hi-C (**Fig 1.15**). On the X chromosome, DCC-mediated domains line up well (**Fig 1.16**), which is reflected in the insulation score profiles (Crane *et al* 2015) - the local maxima and minima for both ARC-C and Crane Hi-C are congruous with domain boundaries in the contact maps (**Fig. 1.17**). In short, ARC-C reproduces large-scale features that were discovered and described in a prior Hi-C map in *C. elegans* mixed embryos.

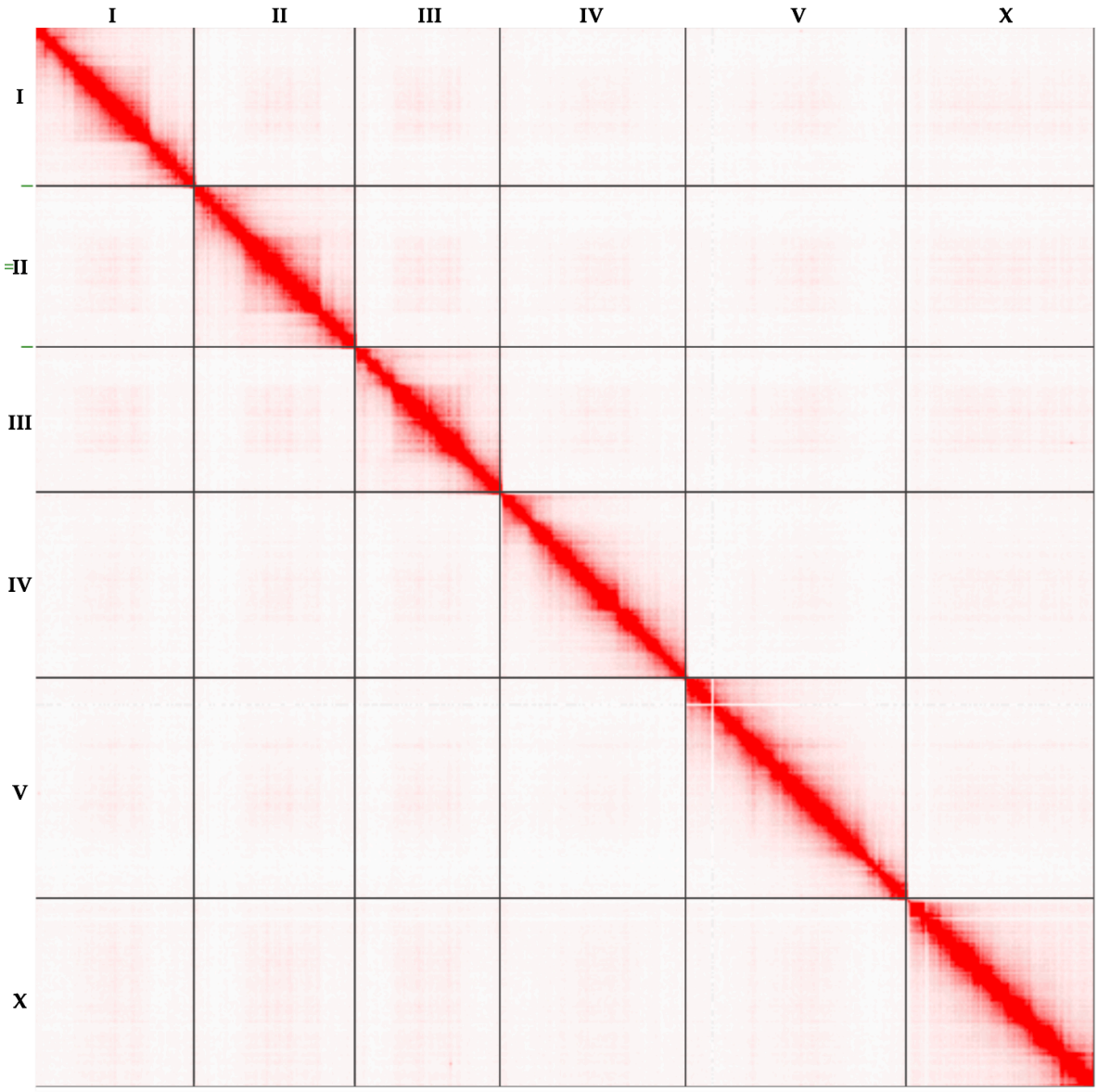


Figure 1.13: Genome-wide intra- and inter-chromosomal contact map for wild-type L3 stage *C. elegans* ARC-C at 50kb resolution.

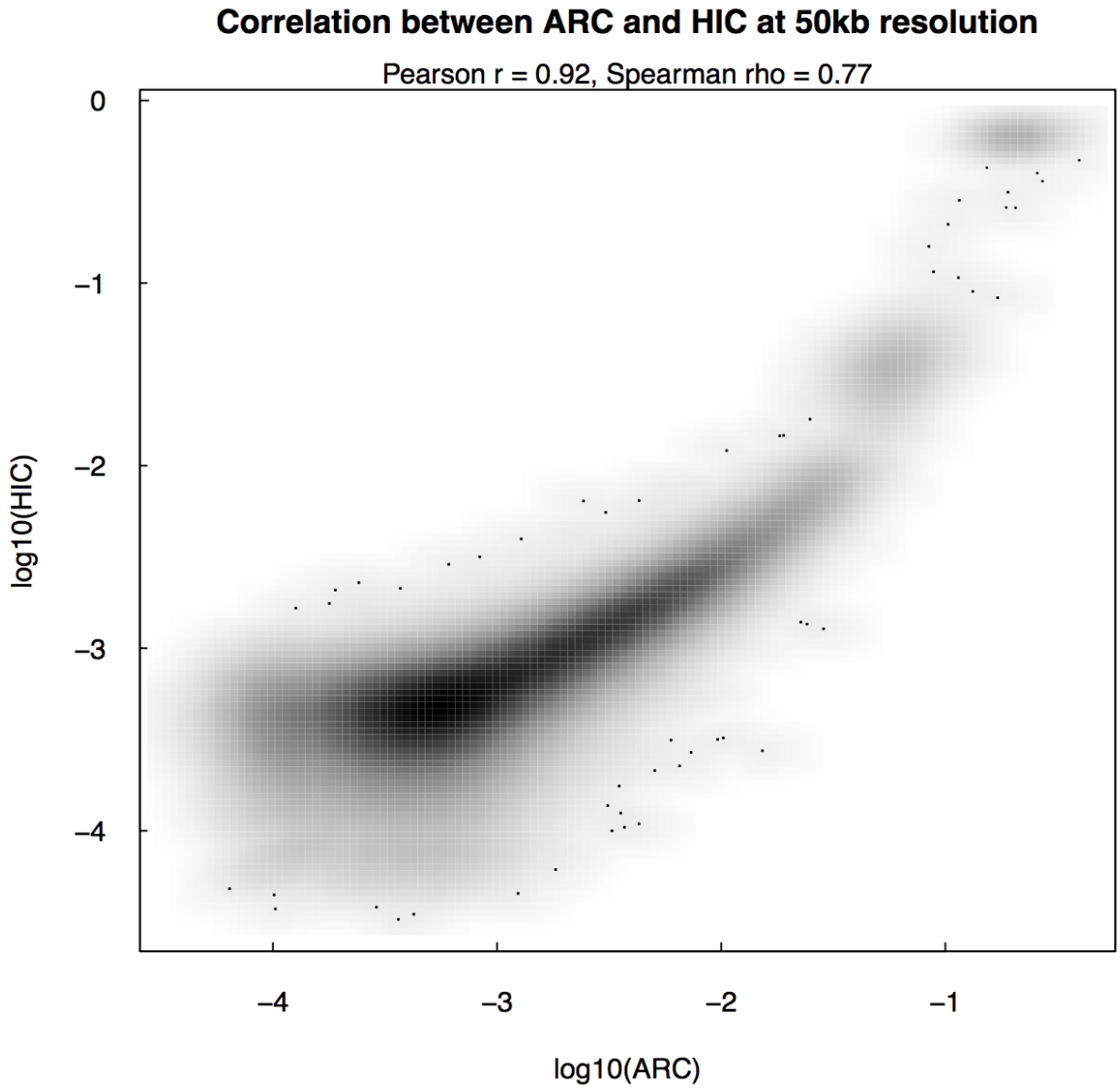


Figure 1.14: Log-log plot of informative ARC-C against informative Crane Hi-C reads binned at 50kb resolution.

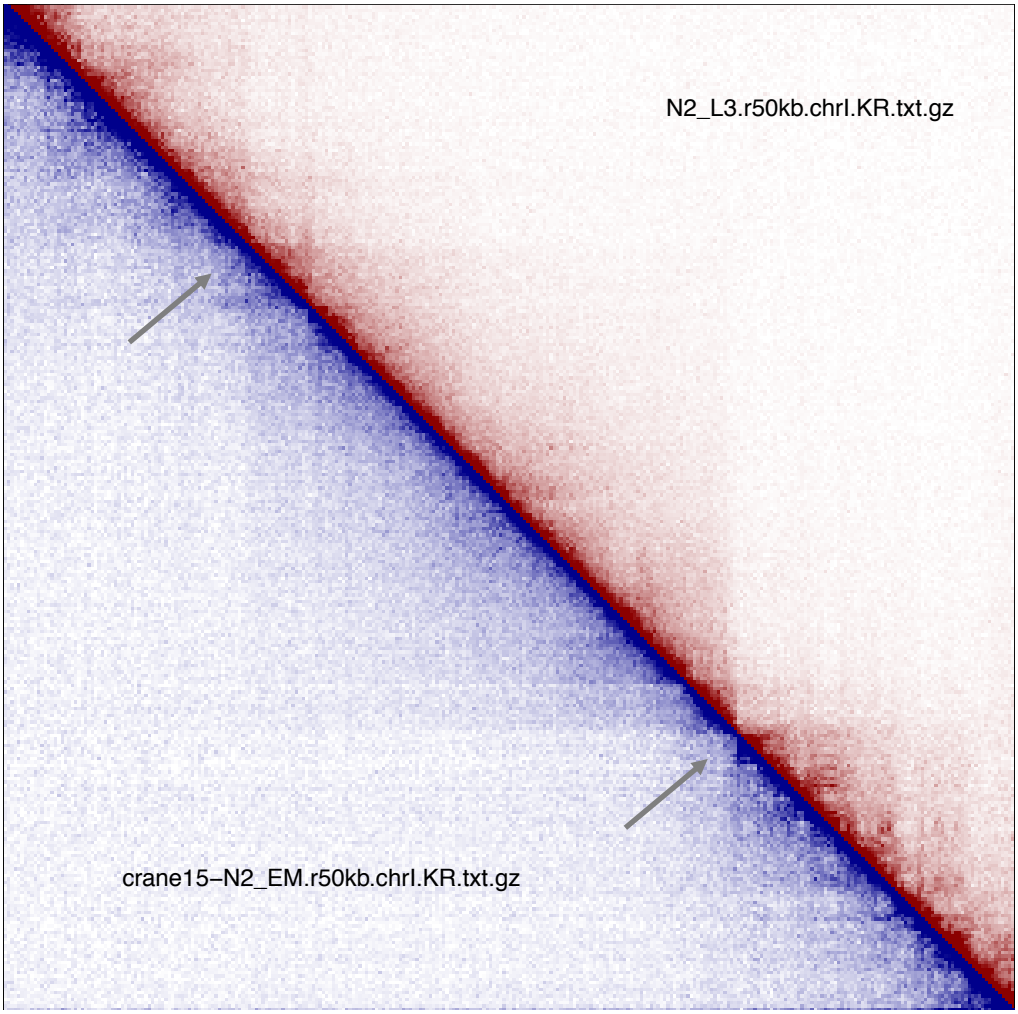


Figure 1.15: Comparison of ARC-C (top) and Crane Hi-C (bottom) contact map for chr I at 50kb resolution. Grey arrows indicate arm-centre transitions.

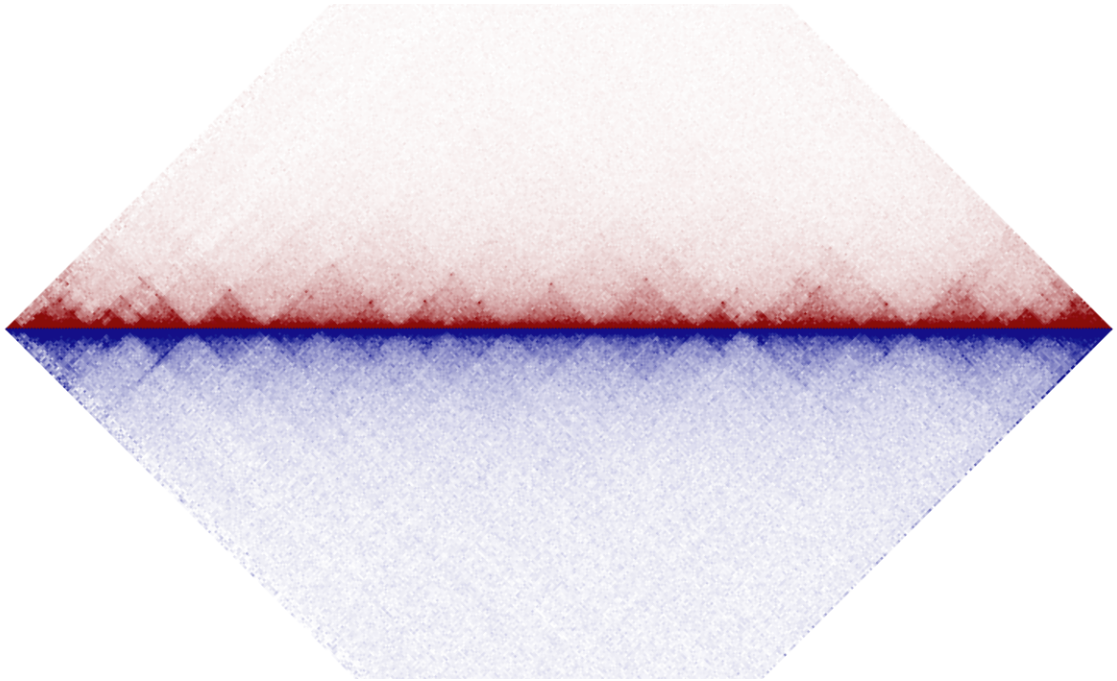


Figure 1.16: Comparison of ARC-C (top) and Crane Hi-C (bottom) contact map for chr X at 50kb resolution.

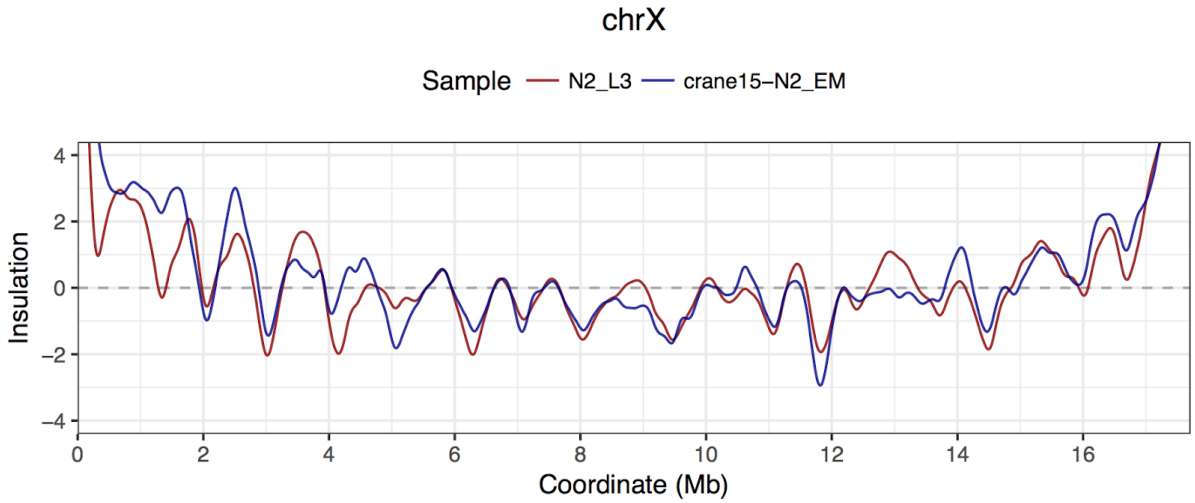


Figure 1.17: Comparison of ARC-C (red) and Crane Hi-C (blue) insulation profiles across chr X at 10kb resolution. An insulation score was calculated which reflects the aggregate of interactions across each bin. The insulation profile is calculated by the log2 ratio of each bin's insulation score and the mean of all insulation scores (Crane *et al* 2015).

To investigate the ability of ARC-C to detect interactions at higher resolution, we browsed informative read pairs in ARC-C and Crane Hi-C through Circos plots without binning. As discussed earlier, *rex* sites are regions where the worm DCC is loaded to down-regulate X chromosome genes by approximately half (Meyer 2010). There are currently 41 *rex* sites and 23 predicted *rex* sites. Crane *et al* (2015) showed that the top 25 *rex* sites, as determined by analyses of DCC components, had statistically significantly higher interaction frequency. This was not qualitatively evident from our Circos plots over a section of the X chromosomes (chr X: 1,200,000 - 1,500,000), which showed a fairly uniform coverage across the genome with no obvious enrichment between any particular loci (each line within the core of the figure plots a single informative read pair; **Fig 1.18**), a similar finding as when we look at single-ended coverage (**Fig 1.8**). In contrast, for ARC-C in the same representative window (chr X: 120,000 - 150,000), there was a qualitative enrichment of interactions between loci corresponding to *rex* sites (each line within the core of the figure plots a single valid read pair; **Fig 1.19**), suggesting an enrichment in interactions between regions of open chromatin.

On the autosomes, we observed interactions between putative regulatory elements, as supported by H3K4me3 (red) and H3K27ac (green) chromatin immunoprecipitation with sequencing (ChIP-seq) binding data (**Fig 1.20**). In the

next section and coming chapter, we develop statistically robust means of identifying enriched interactions and characterise them.

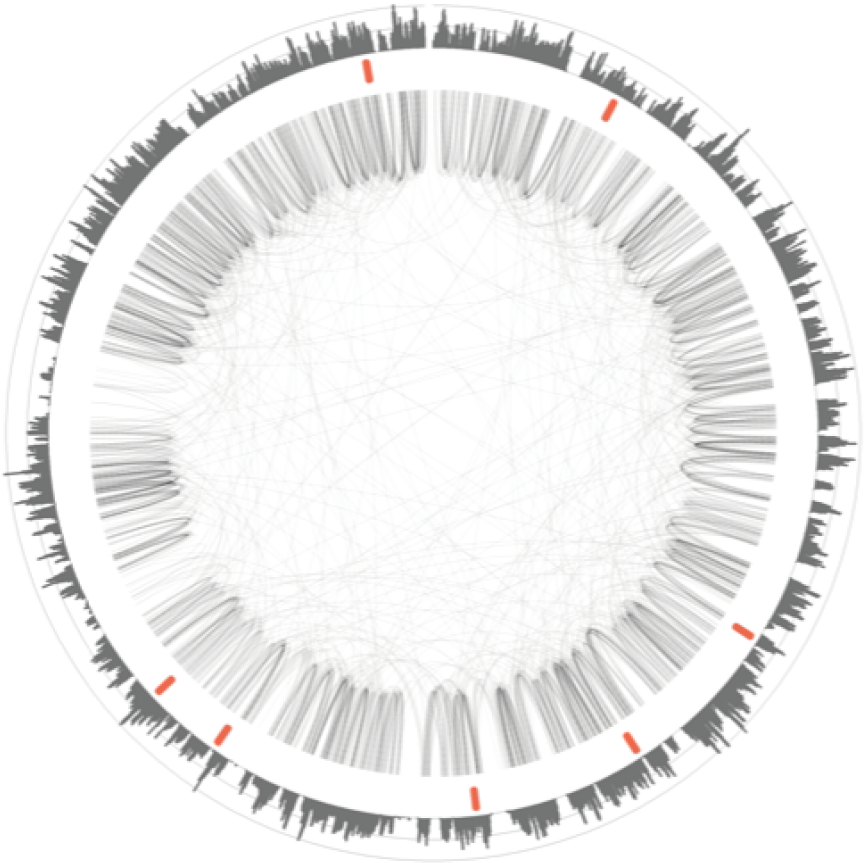


Figure 1.18: Circos plot of HiCUP-processed informative interactions from Crane Hi-C; chrX:1,200,000-1,500,000. Outer to inner core: coverage of informative reads, *rex* sites (red bars), valid interactions (lines).

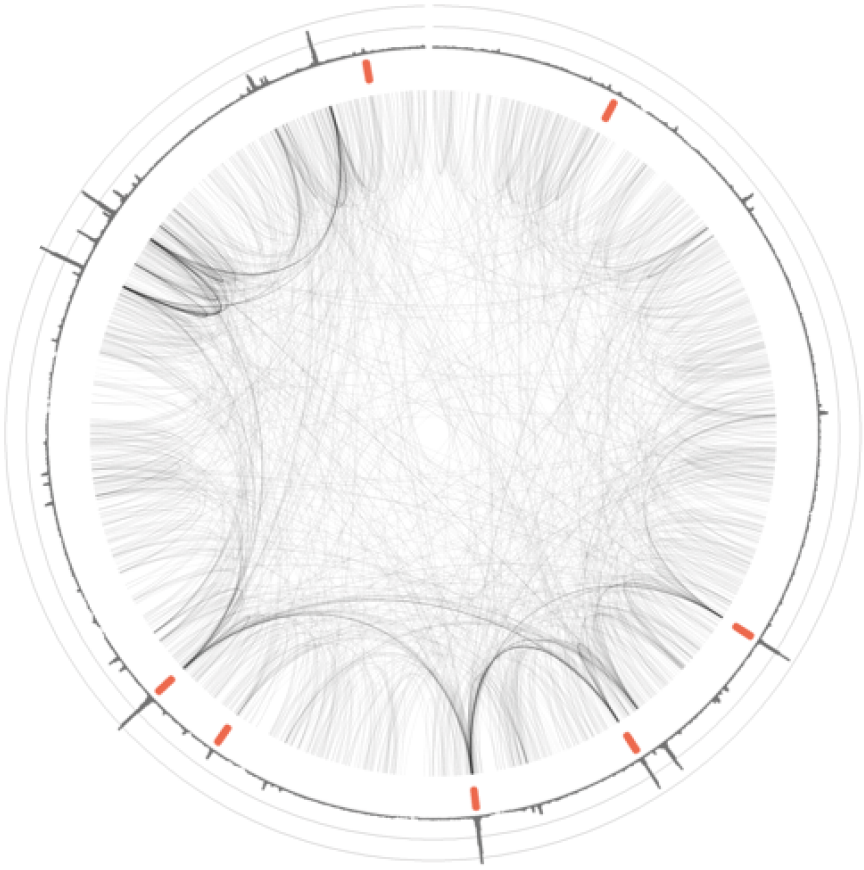


Figure 1.19: Circos plot of informative interactions from ARC-C; chrX: 1,200,000-1,500,000. Outer to inner core: coverage of informative reads, *rex* sites (red bars), valid interactions (lines). Plot shows strong interactions between *rex* sites as alluded to in Crane *et al* 2015.

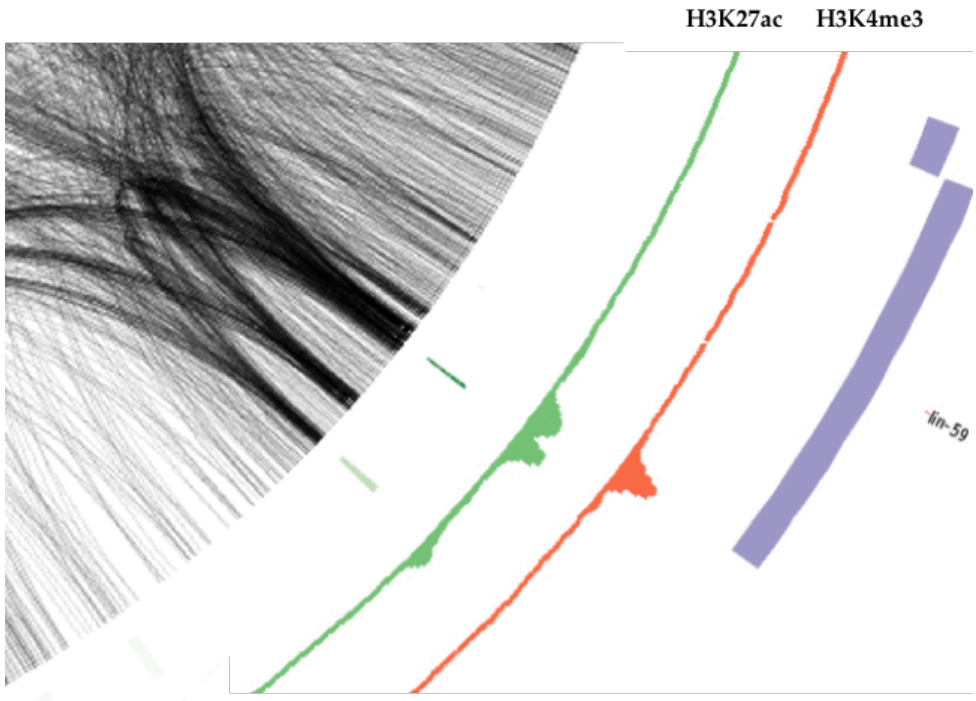


Figure 1.20: Section from Circos plot in chr I: 3,706,000-3,745,000. Outer to inner ring: Refseq genes, H3K4me3 ChIP-seq, H3K27ac ChIP-seq, ATAC-seq peaks, valid interactions.

Calling significant interactions

Biases

To determine significantly enriched interaction loci, we needed to correct technical biases and test observed interactions against an expected background model. ARC-C shares a number of technical biases with Hi-C, such as nucleotide composition, mappability, ligation efficiency, and, to a lesser extent, fragment length (since the fragments generated in ARC-C are less variable in length) (Yaffe & Tanay 2011).

In particular, whilst a difference in accessibility is considered a bias in Hi-C since the genome is supposed to have been digested uniformly, it is less clear the extent to which accessibility differences create false positives in an assay that specifically enriches for accessibility as biological signal. Given the intimate relationship between accessibility and regulatory function - most of the TFs assessed in the ENCODE projects binds exclusively with accessible regions (Thurman *et al* 2012) - it is likely that accessibility reflects the capacity for protein-mediated chromatin interactions.

Indeed, a similar but much less efficient technique with variable accessibility genome-wide - Trac-looping (Lai *et al* 2018) - that used a liberal method for calling significant interactions showed that about 80% of interactions called

corresponded to capture-based or antibody-assisted methods like Hi-ChIP, Capture Hi-C, and ChIA-PET. Moreover, they showed that accessibility and proximity are not sufficient to allow chromatin interactions, arguing against the notion that higher accessibility creates false positives (Lai *et al* 2018). Biological differences in accessibility that is captured in ARC-C could allow for higher sensitivity and specificity in calling significant interactions.

Existing methods for correcting biases

Current correction methods for Hi-C assume that it digests the genome evenly and therefore possesses an even coverage. Technical biases that prejudice this assumption are treated in two main ways: explicitly for known or assumed biases such as fragment mappability, fragment content, and fragment length (Yaffe & Tanay 2011; Hu *et al* 2012: HiCNorm) or implicitly by correcting coverage agnostically as in vanilla coverage normalisation (Lieberman-Aiden *et al* 2009) or matrix balancing (Imakaev *et al* 2012).

Vanilla correction (VC), using the square root of the correction factor used in VC (Sqrtc), and matrix balancing (MB) have been extensively studied in Rao *et al* 2014. Simply, VC corrects for coverage differences by dividing each cell in a contact matrix by the sum of its respective row and then its respective column (Lieberman-Aiden *et al* 2009). VC over-corrects, but can be ameliorated by using

the square root of the correction factor (ie. Sqrtc), producing results that are be similar to more complex solutions (Rao *et al* 2014). Matrix balancing performs the best theoretically and practically as it does not presuppose particular biases and accounts for systemic technical and biological biases (Imakaev *et al* 2012, Rao *et al* 2014, Lajoie *et al* 2015). In essence, it balances the matrix by equalising the sum of every row and column (Lajoie *et al* 2015). That said, all of these methods perform similarly well with Hi-C data - loop calls by HICCUPS, domains, and compartments are correlated and have high overlaps (Rao *et al* 2014) (**Fig 1.21**).

Experimental setup

We experimented with the different Hi-C correction methods (VC, Sqrtc, MB) at 500 bp resolution and used existing packages - Fit-Hi-C (Ay *et al* 2014) and Capture Hi-C Analysis of Genomic Organisation (CHiCAGO) (Cairns *et al* 2016) - that were meant to process probe-based conformation-capture data such as Capture Hi-C or Capture-C, which are more similar to ARC-C, to call significant interactions (**Table 1.23**).

We also tested an off-peak correction method (**Table 1.23**): essentially, this comprises an additional step to take into account differences in background noise due to differences in accessibility. Conceptually, peaks were defined based on the top 10th percentile of *cis* coverage; thereafter, off-peak interactions (i.e. non-peak-

to-peak and non-peak-to-non-peak interactions), which were considered to be largely technical, were taken as a measure of background interactions and used as the bias coefficient to aid in the calling of significant interactions (**Methods**).

Furthermore, instead of using the square root (i.e. 0.5) as in Sqrtc, we investigated the effect of varying the exponent that the coverage is raised to (**Table 1.23** - "Adjusted coverage") on the correction coefficients derived from the correction process, and compared it to the coefficients as derived from matrix balancing (**Fig 1.25**). Lastly, we reduced the contact matrix to encompass only regulatory elements as defined in Jänes *et al* 2018 and performed MB directly on this smaller matrix (**Table 1.23**).

To evaluate the effectiveness of these variations, we performed an Aggregate Peak Analysis (APA) (Rao *et al* 2014, Durand *et al* 2016) of these calls in either ARC-C or Hi-C data. APA is done on an iteratively-corrected contact map and superimposes subsetted windows of the contact map centred on a pair of loci. Loops manifest as a central "dot" in APA and the strength can be quantified with respect to four local neighbourhoods (**Fig 1.22**). We also noted the number of *rex-rex* interactions called, as these represent biologically validated interactions (Crane *et al* 2015), and the overall number of loops called.

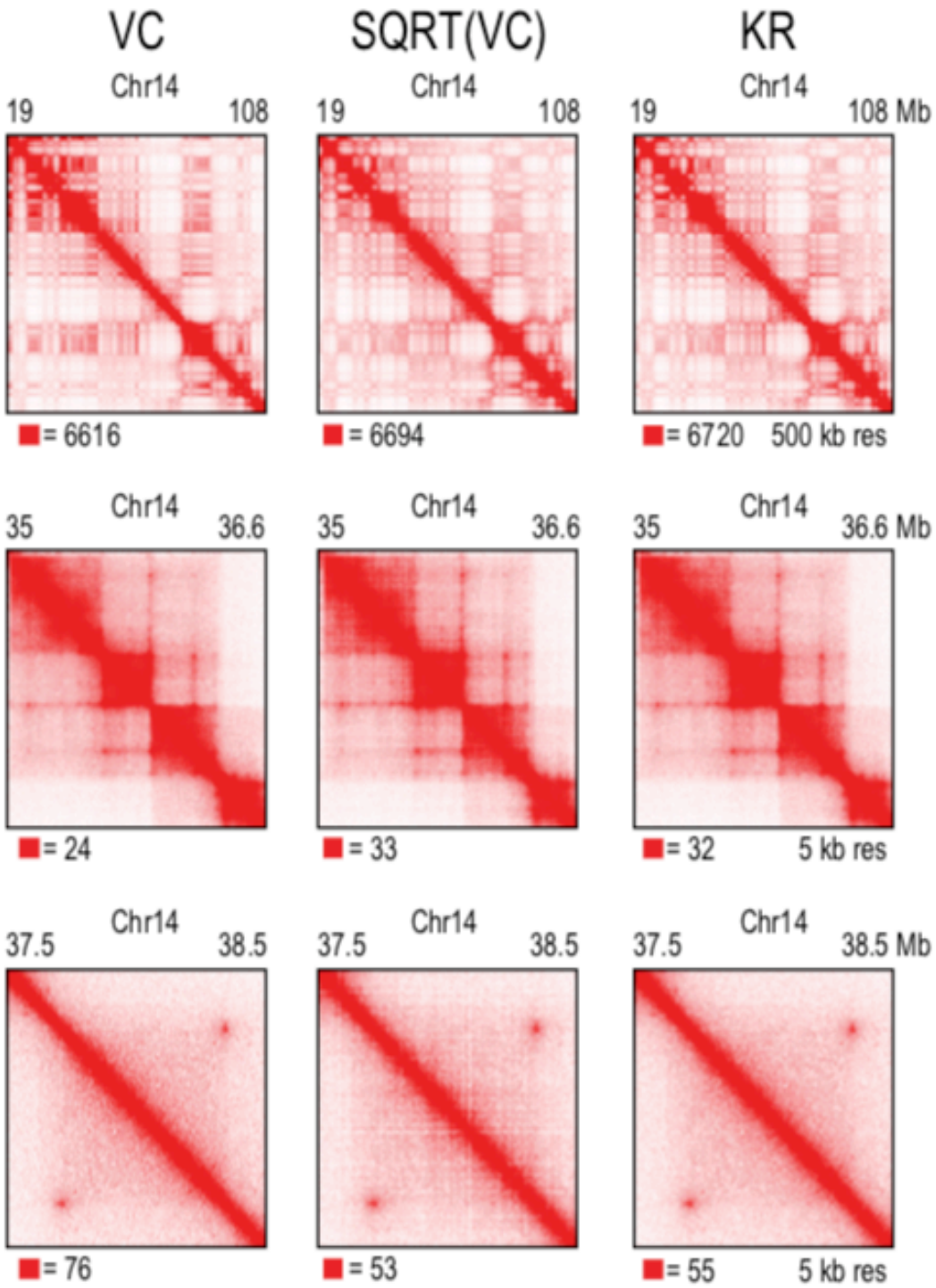


Figure 1.21: Comparison of VC, Sqrtc, and KR (matrix balancing) on GM12878 Hi-C data at different scale (Rao *et al* 2014). With all three methods of implicit normalisation, compartments (top), TADs (middle), and peak foci (bottom) are recapitulated.

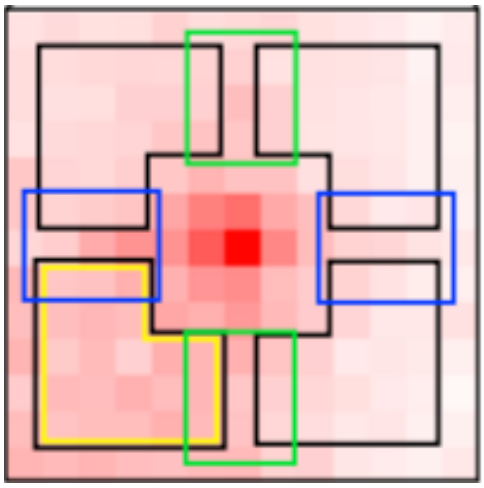


Figure 1.22: Quantifying APA scores or peak foci enrichment. Peaks or "dots" indicate loops whose strength or fold-enrichment can be quantified relative to their local neighbourhoods: black, green, blue, yellow.

The quality of significant interactions was also evaluated by the proportion of loops that were bound by annotated regulatory elements. Briefly, these annotations were based on focal peaks called from ATAC-seq data in multiple developmental stages in *C. elegans*. Combined with transcription initiation and transcriptional elongation signal from capped RNA-seq, peaks were separated into protein-coding promoters, pseudogene promoters, unknown promoters, putative enhancers, non-coding RNA, and others (Jänes *et al* 2018).

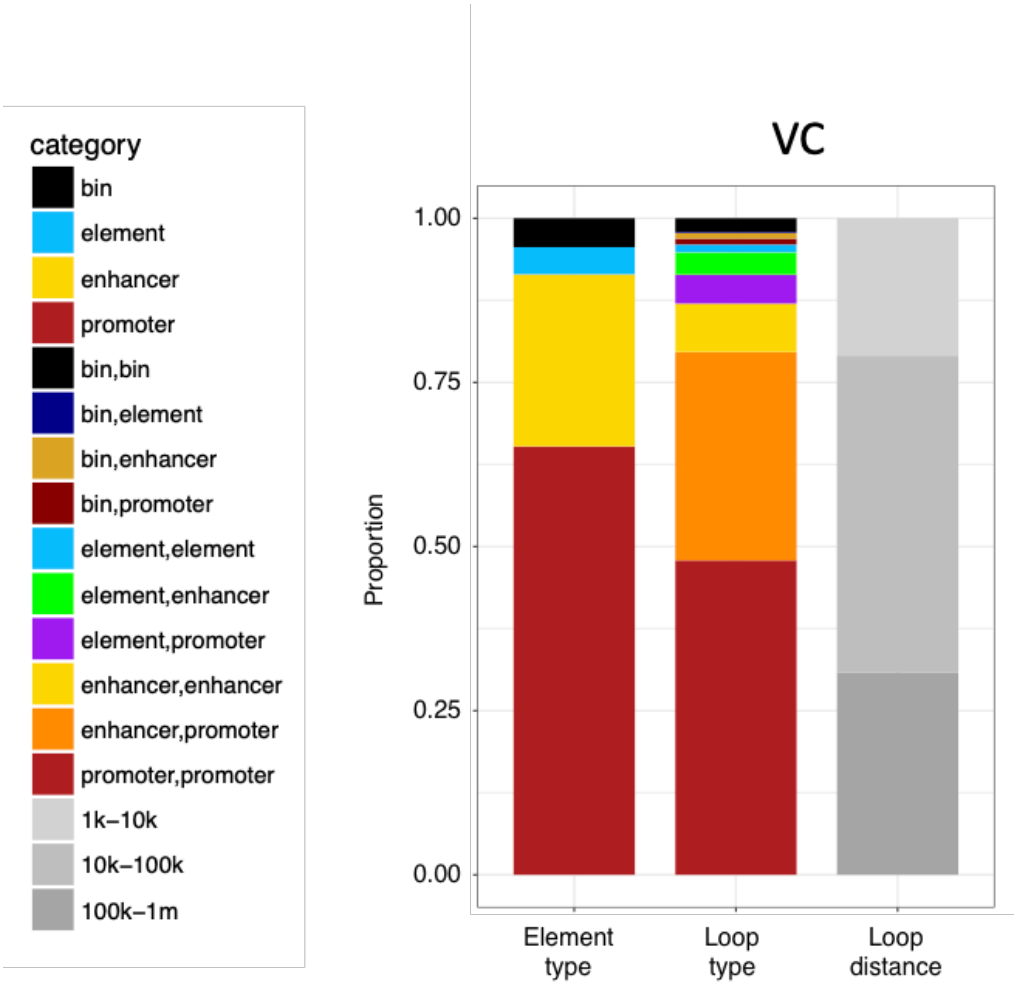
Method	Bins	Off-peak	Add trans	Coverage bias correction	# loops	# loops >20kb	APA ARC-C	APA Hi-C	# *rex-rex*
Fit-Hi-C*; pe1	all			other	106,822	56,416 *(>10kb)			
Vanilla coverage (VC)	all	-	-	coverage	1,255	823	14.78	1.62	43
Square root VC (sqrtc)	all	-	-	coverage^0.5	17,757	9,510	7.70	1.14	73
Binning + Matrix Balancing (MB)	all	-	-	matrix balancing	2,542	1,494	12.66	1.43	54
VC + off peak correction (adapted from Cairns *et al* 2016)	all	+	-	coverage	10,967	5,305	9.19	1.24	63
Sqrtc + off peak correction	all	+	-	coverage^0.5	26,261	12,464	6.94	1.12	73
Adjusted coverage + off peak correction	**all**	**+**	**-**	**coverage^0.87**	**15,014**	**7,410**	**8.38**	**1.16**	**70**
Regulatory elements (RE) MB	RE	-	-	matrix balancing	6,160	332	19.23	1.42	42
CHiCAGO; q0.001	RE	+	+	other	19,721	19,263	12.22	1.08	76

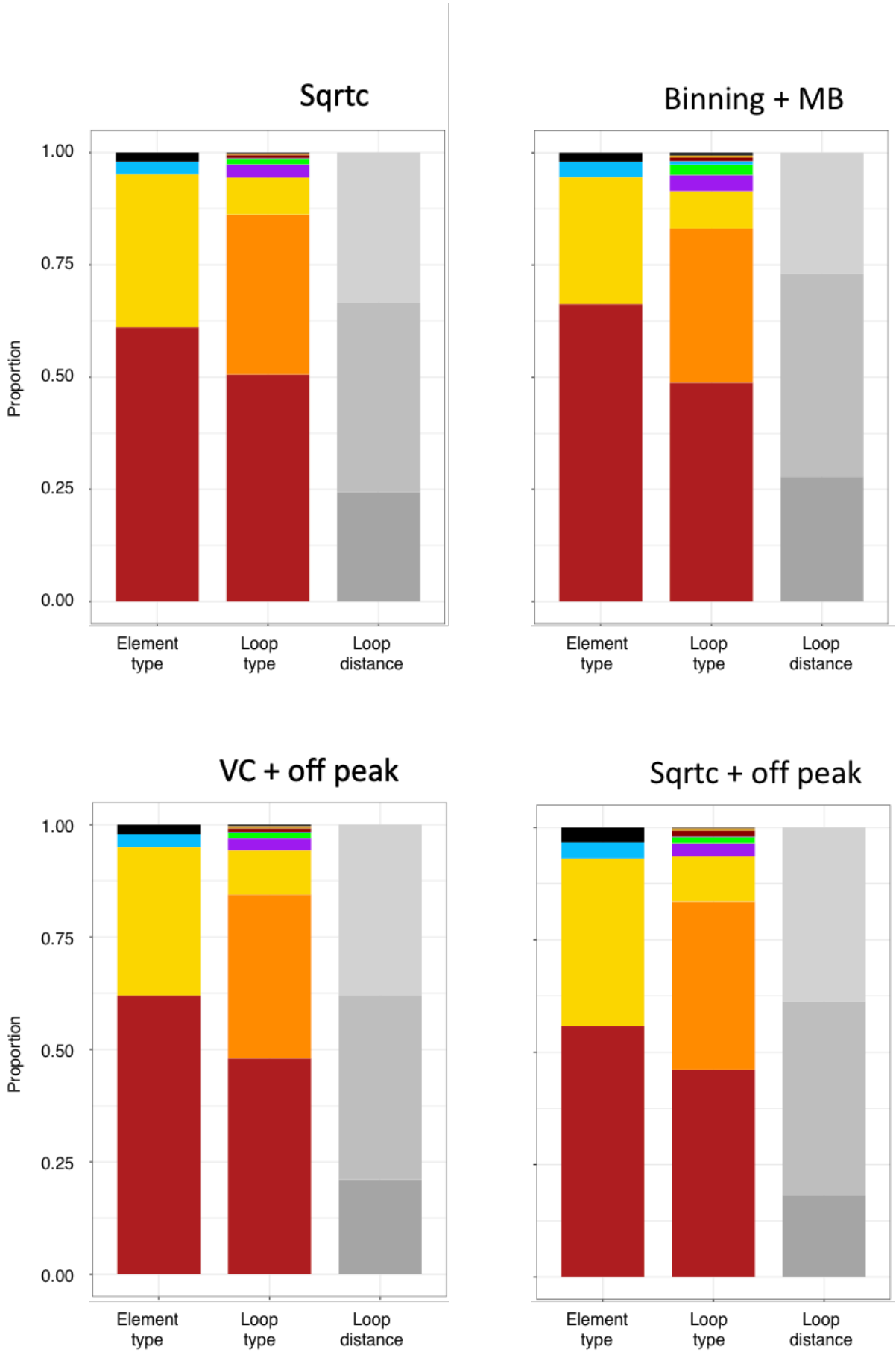
Figure 1.23: Summary of ARC-C normalisation experiments.

Results

All of these methods called significant interactions that covered high proportions of annotated regulatory elements (> 90%) (**Fig 1.24** - left bar: "element type") - and also provided similar proportions of interaction-types - promoter-promoter (middle bar, red), promoter-enhancer (middle bar, orange), enhancer-enhancer (middle bar, yellow), *et cetera* (**Fig 1.24**), indicating that ARC-C is robust and enriches for interactions at and between regulatory elements

However, the distance distribution for significant interactions appears to be sensitive to the method used. Performing matrix balancing on a matrix consisting of regulatory elements resulted in mostly short distance interactions (**Fig 1.24** - RE + MB), whilst CHiCAGO only produced medium-long range interactions (**Fig 1.24** - CHiCAGO). The other methods support a distance distribution in which the majority were medium range (10 - 100kb), consistent with observations that promoter-enhancer interactions typically occur within 200kb (Ma *et al* 2015, Tang *et al* 2015, Mumbach *et al* 2017, Cao *et al* 2017).


category
bin
element
enhancer
promoter
bin,bin
bin,element
bin,enhancer
bin,promoter
element,element
element,enhancer
element,promoter
enhancer,enhancer
enhancer,promoter
promoter,promoter
1k–10k
10k–100k
100k–1m
VC
Proportion
1.00
0.75
0.50
0.25
0.00
Element type
Loop type
Loop distance


Sqrtc
Binning + MB
VC + off peak
Sqrtc + off peak
Proportion
1.00
0.75
0.50
0.25
0.00
Element type
Loop type
Loop distance

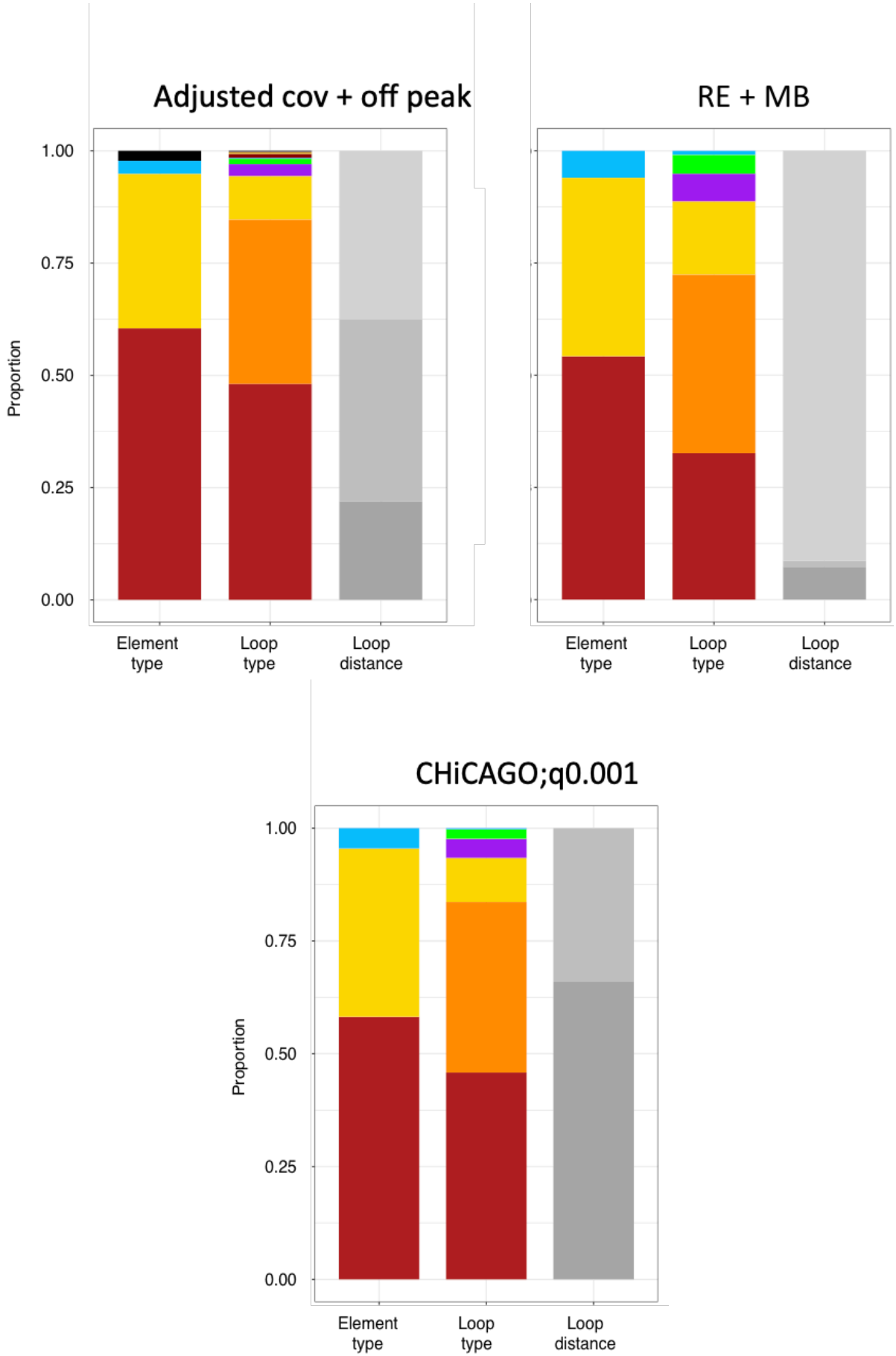

Adjusted cov + off peak
RE + MB
CHiCAGO;q0.001
Proportion
1.00
0.75
0.50
0.25
0.00
Element type
Loop type
Loop distance

Figure 1.24: Breakdown of significant interactions by interaction-types and distances.
Left bar: promoters (red), enhancers (yellow), unannotated elements (blue), non-accessible (black).
Middle bar: promoter-promoter (red), enhancer-promoter (orange), enhancer-enhancer (yellow), purple (promoter-unannotated), green (enhancer-unannotated), blue (unannotated-unannotated).
Right bar: 100kb - 1Mb (dark grey), 10kb - 100kb (medium grey), 1kb-10kb (light grey).

In terms of the number of loops, VC, genome-wide matrix balancing, and regulatory elements matrix balancing have the fewest number of called interactions - 1255, 2542, and 6160 respectively (**Table 1.23**). These methods tend to over-correct and, accordingly, they produce the lowest number of biologically validated *rex-rex* interactions - 43, 54, and 42 respectively (**Table 1.23**).

The process of matrix balancing results in a correction value for each column (or row) in the given matrix. We compared these values against ones we obtained when adjusting the exponents (0.5, 0.6, 0.7, 0.87, 1) during coverage normalisation "0.87" gave similar correction values as matrix balancing (**Fig 1.25**), suggesting that that is the appropriate exponent to select. This method of adjusted coverage produced a high APA score (8.38) in ARC-C and captured a large number of *rex-rex* interactions (70) (**Table 1.23**).

Prima facie, CHiCAGO seems reasonable (**Table 1.23**), calling a reasonable number of significant interactions (19,721), having a high APA score (12.22), and a

large number of *rex-rex* interactions (76), but it is limited because it does not call short range interactions (<10kb) and might miss intragenic interactions (median gene length in *C. elegans* = 1,956bp).

We selected the method that combined an "adjusted coverage normalisation" and "off-peak correction" (bold in **Table 1.23**). It is the most theoretically robust normalisation method we have that corrects for coverage and accounts for accessibility-associated technical noise. It balances theoretical rigour, a high APA score in ARC-C, and a large number of validated *rex-rex* interactions. Approximately 12 million *cis*, informative reads yielded 15,014 significant interactions (**Table 1.23**).

In retrospect, a more suitable way of validating these methods would be to test them functionally. This would involve splitting call-sets by confidence levels, randomly sampling pairs of loci within each interval and conducting 3D-FISH to test if they were in closer proximity with each other as opposed to random pairs of loci that are of the same 2D distance apart.

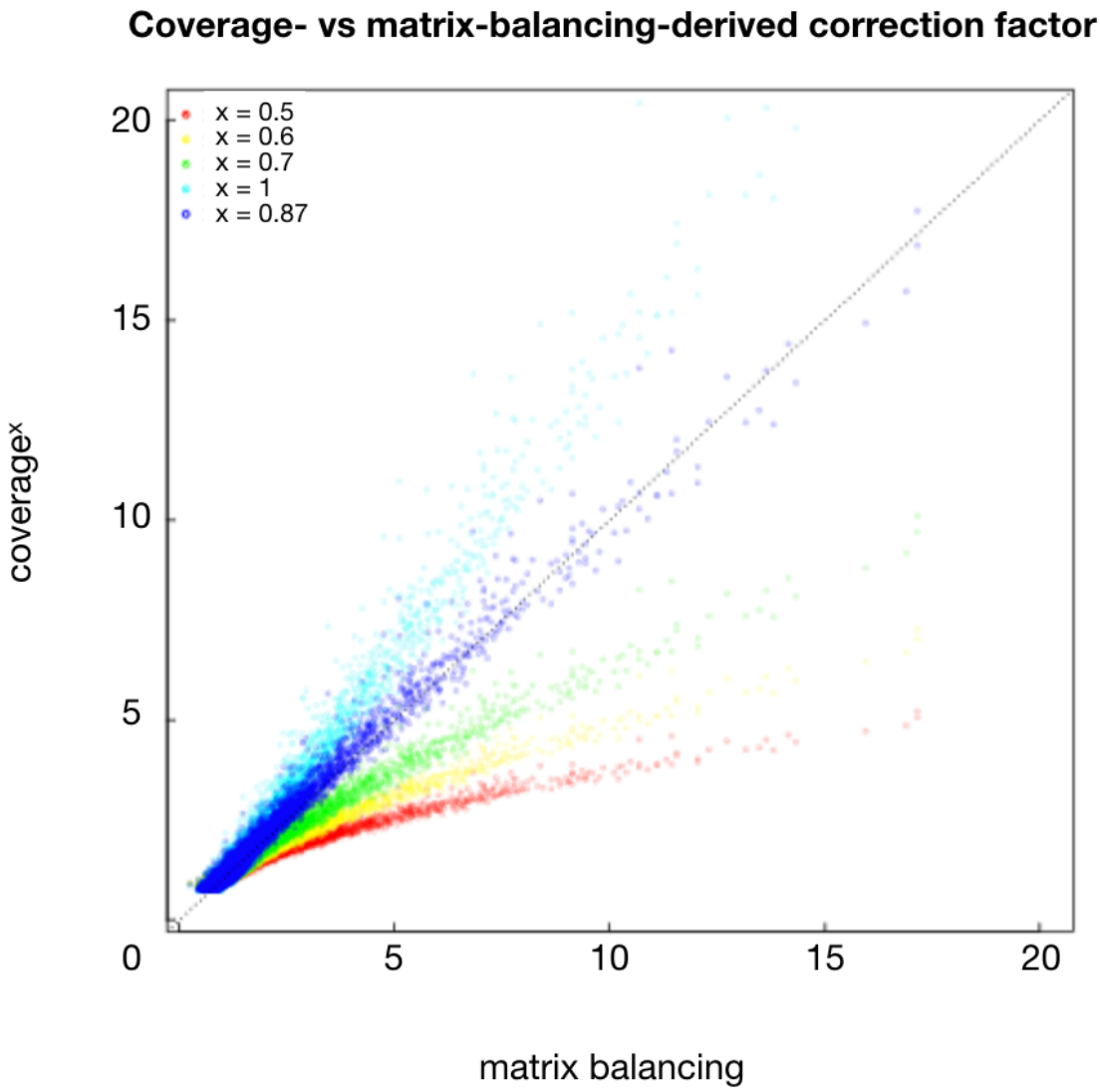


Figure 1.25: Plot of correction coefficients derived from matrix balancing against various adjusted coverages at 500bp resolution.

Comparison with DNase Hi-C

ARC-C was developed independently of DNase Hi-C (Ma *et al* 2015), which also aimed to overcome the restriction fragment resolution limit by the use of DNase I. *In situ* DNase Hi-C was later developed as an improved version (Ramani *et al* 2016, Ma *et al* 2018) of DNase Hi-C; proximity ligation was performed within

intact nuclei instead of in solidified agarose gels and resulted in better *cis*-to-*trans* ratios (Deng *et al* 2015, Ramani *et al* 2016). DNase Hi-C libraries can be subsetted with the use of DNA probes ("targeted DNase Hi-C"; Ma *et al* 2015) to capture regions of interest, analogous to Capture Hi-C and Capture-C.

DNase Hi-C and ARC-C apply the same endonuclease to fragment chromatin in nucleus, but the similarity ends there. Importantly, DNase Hi-C digests chromatin to an extent that abrogates an enrichment of cuts at DHS: the ideal digestion for DNase Hi-C would be when most of the fragments are below 1kb (Ma *et al* 2018). Accordingly, DNase Hi-C has a similar accessibility bias as restriction enzyme-based Hi-C at both local and large scales, and slightly outperforms it in terms of GC content, mappability, and the percentage of genome covered (Ma *et al* 2015). As a result, coverage of reads in DNase Hi-C is fairly even across the genome and at DHSs (**Fig 1.26**).

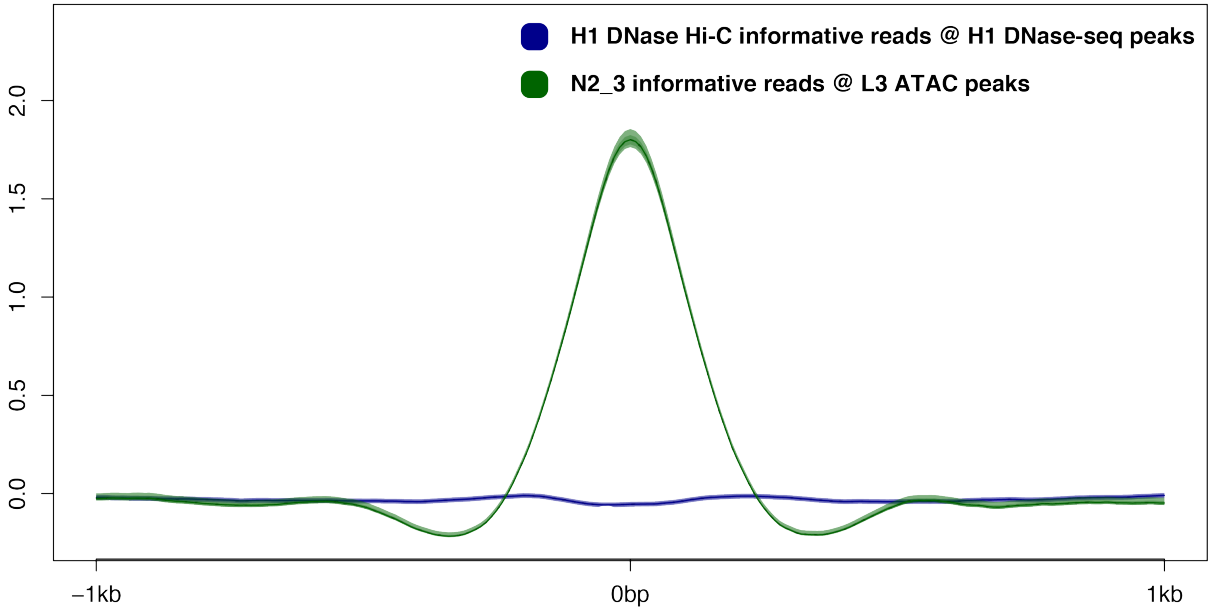


Figure 1.26: Informative coverage in DNase Hi-C (blue) and ARC-C (green) centred around their corresponding hypersensitive sites.

I endeavoured to compare the performance of targeted DNase Hi-C *vis-a-vis* ARC-C. However, there are important points of divergences since ARC-C was done in *C. elegans*, whilst targeted DNase Hi-C was performed in much larger mouse and human genomes. Moreover, relatively few loci were selected (113-1,001) for targeted DNase Hi-C. Comparisons made here are done with these caveats in mind.

To compare the performance of DNase Hi-C *vis-a-vis* ARC-C, I looked at the several indicators where applicable (**Fig 1.27**): *cis*-to-*trans* ratio, the informative efficiency (i.e. *cis* reads mapped farther than 1kb away), percentage of *cis*-

informative reads-on-target (i.e. at least one end of a read pair within a target region, with the other being at the rest of the genome and not within the same target region), significant interactions called, their resolution, and overlap with DHSs.

	targeted DNase Hi-C: H1-ESC lincRNA	ARC-C: N2 L3
***cis* ratio (%)**	62.82	78.07
***cis* informative efficiency (%)**	31.58	7.77
reads-on-target (%)	25.49	43.70
***cis* informative read pairs**	35,670,284	12,192,654
significant interactions	12,739	15,014
resolution	1kb	500bp
overlap with DHS (%)	14.33	98.21

Figure 1.27: Summary of statistics in targeted DNase Hi-C and ARC-C.

Of note, beyond sequencing through endogenous DNA fragments into the bridge adaptors, there is no way to distinguish informative read pairs in DNase Hi-C since cut sites are relatively random and there is no strict size selection before sequencing. I took a 1kb cutoff as "informative" to have a basis for

comparison, since 1kb was used in ARC-C and is the ideal size for DNase I digestion in DNase Hi-C. In this case, ARC-C has less informative efficiency (i.e. % of *cis* informative/valid deduplicated reads), which is unsurprisingly as it lacks a biotin-street avidin enrichment step. As mentioned earlier, this is being worked on in the next iteration of the method.

Importantly, when it comes to the reads-on-target (% of *cis* informative read pairs with at least one at a target region and one at the rest of the genome/all *cis* informative read pairs), it appears that a typically low capture efficiency and preponderance of read pairs mapping to the same target region meant that, overall, DNase Hi-C has a lower percentage of reads-on-target than ARC-C. In addition, whilst not attempted yet, ARC-C libraries can also be subjected to a similar capture protocol.

Also, capture libraries are essentially subsets of the original libraries, and possess the same biases and background noise. Despite the use of Fit-Hi-C on approximately 36 million read pairs, only 12,739 significant interactions were called. Comparatively, the use of Fit-Hi-C in ARC-C produced 106,822 significant interactions (**Fig 1.23**) and our conservative caller produced 15,012 significant interactions from about 12 million read pairs. This may be attributed by a higher amount of background non-bait-to-non-bait read pairs in targeted DNase Hi-C

libraries. However, the use of different genomes with different chromatin organisation (and thus, different levels of distance-mediated decay of contact frequency) and different intergenic distances makes such a comparison incomplete.

Accordingly, much fewer significant interactions in targeted DNase Hi-C showed regulatory potential - as defined by their overlap with at least one DHS (**Fig 1.27**). The overlap for targeted DNase Hi-C was done with ENCODE-curated DNase-seq peaks in H1 cells (GEO: GSM736582). Chromatin accessibility was taken as a proxy for regulatory potential (reviewed in Klemm *et al* 2018) - a majority of assayed TFs bind open chromatin almost exclusively (Thurman *et al* 2012). Moreover, in *C. elegans*, distal ATAC-seq peaks (>1 kb from TSS) showed enhancer activity in transgenic assays, supporting the use of accessibility as a proxy for regulatory potential. The difference in overlap with DHS between DNase Hi-C and ARC-C validates ARC-C's power in calling regulatory interactions.

Whilst DNase Hi-C seems to be similar to ARC-C, they are fundamentally different techniques with different research goals. As it is, ARC-C is more suited for a multi-scale study of genome organisation that includes an interrogation of regulatory interactions.

CHAPTER II: USING ARC-C TO DEFINE CHROMATIN INTERACTIONS AT HIGH RESOLUTION

A high resolution view of the regulatory landscape is necessary for a better understanding of the genetic elements and factors underlying gene regulation. With ARC-C, we could call about 15,000 significant interactions in wild-type L3 stage worms. In this chapter, I sought to further characterise these interactions.

Regulatory interactions

As discussed in **Chapter I**, based on annotations from Jänes *et al* (2018), we observed 98.21% of significant interactions between regulatory elements. Promoter-promoter interactions (PP) took up 29.67% of all significant interactions; 44.31% were promoter-enhancer interactions (PE), and 24.23% were enhancer-enhancer interactions (EE). Of the total number of unique interacting elements (13,457), 56.9% were annotated promoters, while 36.7% were putative enhancers.

Promoter-promoter interactions

We wanted to assess the effect of distance on the distribution of each interaction type. For that, we separated all significant interactions into 6 distance intervals, each containing the same number of interactions. The expected proportion for each interaction type was simulated based on a permutation of all enhancers and promoters within particular distance intervals. We then calculated the overall enrichment of observed interaction types over the expected values (**Fig 2.1A** & **Fig 2.1B**). *Prima facie*, an increasing distance should enrich for interactions that are not proximity-driven; that is, the regulation and mediation of long-distance interactions presumably require specific factor-mediated mechanisms (Nolis *et al* 2009; Sanyal *et al* 2012).

Intriguingly, while the proportion of PP increased over greater distances, that of EE followed a decreasing trend (**Fig 2.1A** & **Fig 2.1B**), suggesting that the mechanism for PP interactions is predominantly actively driven while EE interactions are limited by physical distance. PE interactions remain fairly stable but follow a slightly decreasing trend as distances increase (**Fig 2.1A** & **Fig 2.1B**), which could be interpreted as an equilibrium between both forces. In fact, a multi-organismic (*C. elegans* included) meta-analysis proposed that most PE pairs are indiscriminately compatible with each other and are governed by proximity than specific interactions (Quintero-Cadena & Sternberg *et al* 2016). In mice, the

activation of an enhancer by fibroblast growth factor "ripples" to neighbouring genes, resulting in their up-regulation (Ebisuya *et al* 2008). These trends are accordant with the model where strong loops (i.e. CTCF-cohesin mediated loops in other organisms and presumably PP interactions in *C. elegans*) bring other regulatory elements in close proximity within permissive chromatin compartments for gene regulation (Ren *et al* 2017, Isoda *et al* 2017). Incidentally, the largest drop in EE proportion occurs from the 10 to 20kb (**Fig 2.1A** & **Fig 2.1B**), around the median length of active chromatin state domains (19,500 bp) in *C. elegans*.

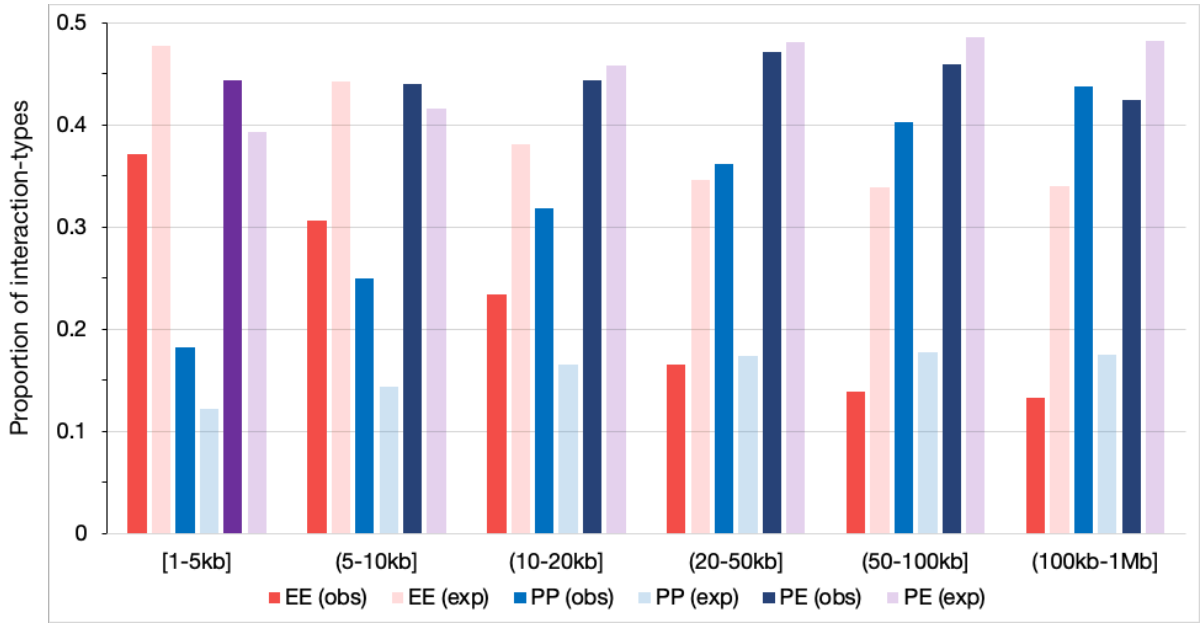


Figure 2.1A: Proportions of EE (red), PP (blue), and PE (purple) interactions at different distance intervals (1-5kb, 5-10kb, 10-20kb, 20-50kb, 50-100kb, 100kb-1Mb). The observed proportions (solid) were contrasted with expected proportions (transparent) based on random permutations of enhancers and promoters within distance intervals.

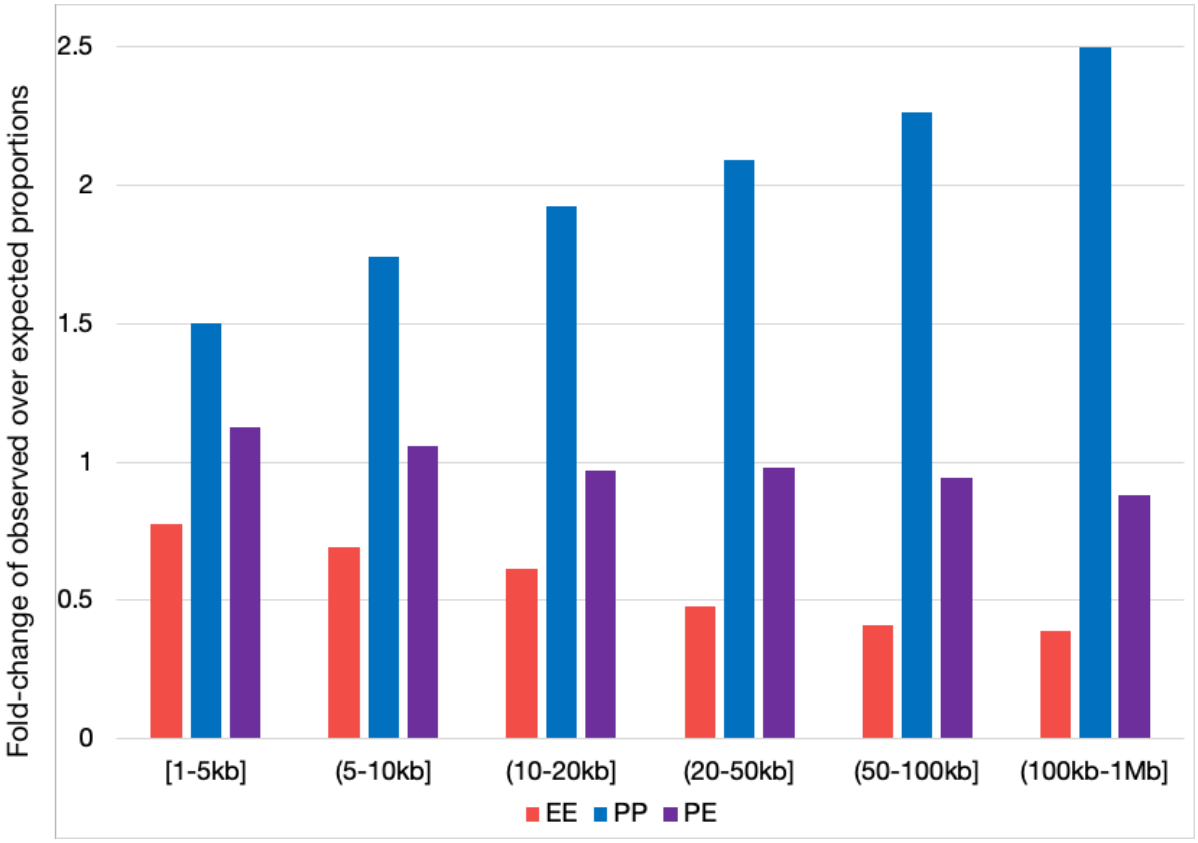


Figure 2.1B: The fold-change of observed over expected proportions for each interaction-type at different distance intervals (1-5kb, 5-10kb, 10-20kb, 20-50kb, 50-100kb, 100kb-1Mb).

Multiple studies have shown that interacting PP pairs are transcriptionally co-regulated - spatially or temporally (Mercer & Mattick 2013, Schoenfelder *et al* 2015, Ibn-Salem *et al* 2016, Soler-Oliva *et al* 2017, Belyaeva *et al* 2017). We wanted to assess if this occurred in *C. elegans* as well. To do so, we first measured the expression correlation between each gene for our PP pairs using gene expression data in different developmental stages from modENCODE (Gerstein *et al* 2010).

We next looked at the relationship between distance and expression correlation. Distance between genes of unique PP pairs were segmented into

quintiles that contained approximately equal number of genes (0-6kb, 6-16kb, 16-36kb, 36-110kb, 110kb-1Mb). Within these defined intervals, gene pairs were randomly shuffled and expression correlation for these shuffled pairs was re-calculated. At all distance intervals, the observed expression correlation of PP pairs was higher than by chance, suggesting the connections were functional (**Fig 2.2**). The general downtrend in expression correlation with increasing distance for randomly shuffled gene pairs agrees with a study that looked at expression correlation between genes and their 100 nearest neighbours; they found that expression correlation decays exponentially with distance until approximately 10 to 20 kb and took it as evidence for enhancer sharing and proximity guided chromatin interactions (Quintero-Cadena & Sternberg 2016). However, our observed pairing appears to buck this trend. The difference between observation and expectation is higher with increasing distance (from 0.036 for the first quintile to 0.080 for the last quintile) (**Fig 2.2**). High correlations at shorter distances (0-16kb) likely represent the co-expression of clustered genes that has been observed in worms (Lercher *et al* 2003), humans (Lercher *et al* 2002) and flies (Spellman & Rubin 2002). That the difference between observed and expected is high at long distances suggests that long distance interactions are also likely to be meaningful.

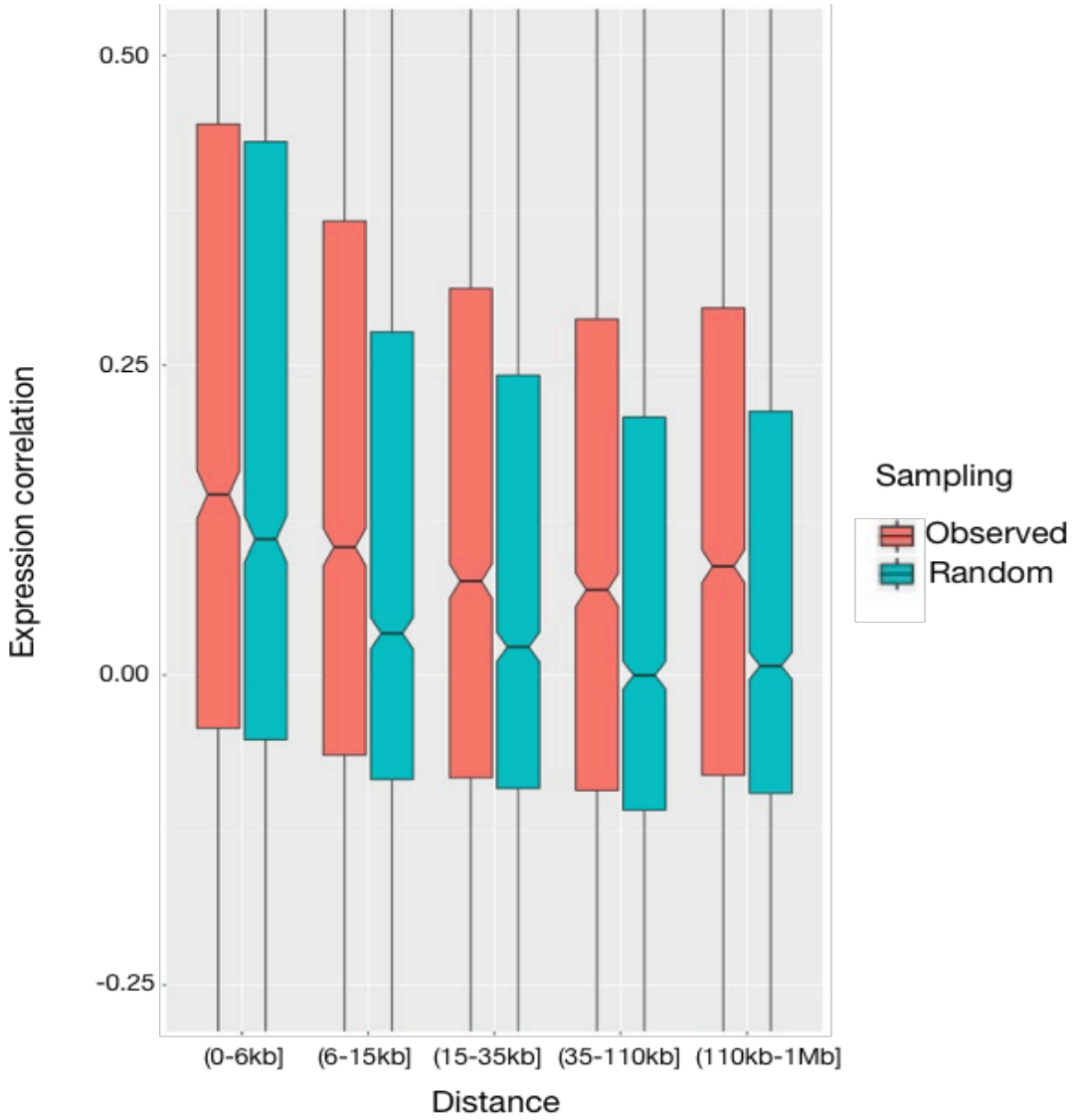


Figure 2.2: Expression correlation between promoter-promoter pairs at different distance intervals (0-6kb, 6-16kb, 16-36kb, 36-110kb, 110kb-1Mb). Observed (red) and randomly shuffled or expected (green) expression correlation are paired for each distance interval.

High occupancy target (HOT) regions are unique genetic elements that have an inordinately large number of factors binding. In the thesis, we define HOT regions as loci that have more than 29 factors binding (top 10^{th} percentile) in L3 stage larvae. The large number seems antithetical to contemporaneous transcription factor binding. HOT regions lack sequence-specific binding motifs

(Gerstein *et al* 2010), which calls into question the mechanism of recruitment for many of these factors, but are enriched for features of chromatin activity. They are CpG-dense promoters (Chen *et al* 2014), near ubiquitous genes (Yip *et al* 2012), have high accessibility and nucleosome turnover (The modENCODE Project Consortium 2010), and have high overlaps with DCC-mediated domain boundaries (87.5%, $p < 0.001$; Crane *et al* 2015). Furthermore, in humans, the presence of TF binding at HOT regions is a strong predictor of RNA polymerase II recruitment and transcriptional activity (Foley & Sidow 2013).

Since the vast majority of modENCODE and in-house ChIP-seqs were performed in whole animals, we cannot exclude the possibility that different factors bind at the same location in different cell-types. However, there is evidence from L3 stage sorted worm myocytes showing that a large number of unrelated factors do bind to HOT regions (data not shown), as well as parallels in *Drosophila* cell lines (Gerstein *et al* 2010).

We sought to determine if they had a role in organising promoter-promoter interactions. Genes were classed as 'HOT' if any of their annotated promoters overlapped HOT regions. Gene expression correlation of PP gene pairs was then compared to randomly shuffled pairs that were controlled for having similar numbers in each category (i.e. HOT-HOT, HOT-nonHOT, nonHOT-nonHOT).

Overall, PP pairs had higher expression correlation than by chance (**Fig 2.3**, $p < 0.001$). However, when segregated by categories, we found that the difference between observation and expectation is driven exclusively by the nonHOT-nonHOT category (**Fig 2.3**). The nonHOT genes in this group have a median CV of 2.112 (around 50^{th} percentile of all genes), indicating tissue-specificity or conditional expression, while HOT genes have a median CV of 0.707 (around 15^{th} percentile of all genes), indicating ubiquitous or broad expression. Accordingly, HOT-HOT PP interactions are enriched for being within and between active chromatin state domains (Fisher's exact test, $p < 0.001$) and nonHOT-nonHOT interactions are frequently found within regulated domains ($p < 0.005$). In all, HOT regions are associated with active domains and mediate interactions nonspecifically, while within regulated domains, nonHOT genes are specifically paired.

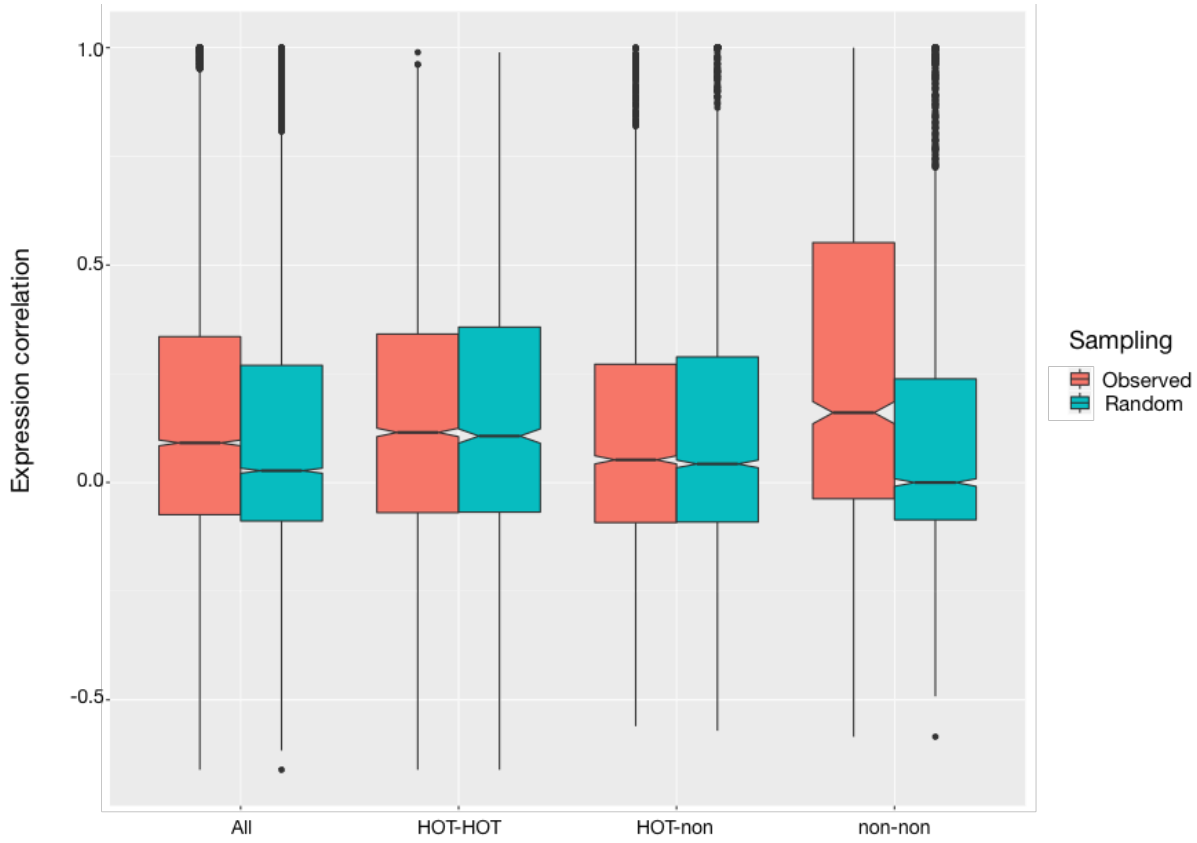


Figure 2.3: Expression correlation of promoter-promoter pairs for All, HOT-HOT, HOT-nonHOT, and nonHOT-nonHOT classes. Observed (red) and randomly shuffled or expected (green) expression correlation are paired for each distance interval.

Promoter-enhancer interactions

When we turn to PE pairs, we find evidence for enhancer additivity in *C. elegans*. In mouse embryonic stem cells, enhancers interact with only 1.32 promoters on average, while promoters interacted with 5.97 putative enhancers on average (Sahlén *et al* 2015), which suggests enhancer sharing by promoters is not frequent and that multi-gene interaction hubs are not extensive. In *C. elegans*, promoters interact with 3.31 enhancers on average. In addition, gene expression

positively correlates with the number of interacting enhancers (**Fig 2.4**), supporting enhancer additivity, which is consistent with transgenic assays where ectopically introduced enhancers boosted expression of nearby native genes in *C. elegans* (Quintero-Cadena & Sternberg 2016).

The number of enhancers per gene is also inversely related to gene expression CV: the promoters of tissue-specific genes have fewer connections to enhancers than those of ubiquitously expressed genes (**Fig 2.4**). This contrasts with literature that claims conditionally-regulated genes require enhancers to fine-tune gene expression as opposed to housekeeping or ubiquitous genes that have simpler regulatory structures (discussed in Farré *et al* 2007). Likely, enhancer additivity is more prevalent than enhancer redundancy in *C. elegans* gene regulation.

There is a caveat here: with ATAC-seq, we detected fewer promoters for genes with higher CV (Jänes *et al* 2018) (**Fig 2.5**), and will likely underestimate the number of weakly accessible enhancers due to a lower sensitivity in detecting tissue-specific elements in a heterogeneous, whole-animal sample. This also applies to ARC-C and might lead to fewer observed interactions for tissue-specific promoters.

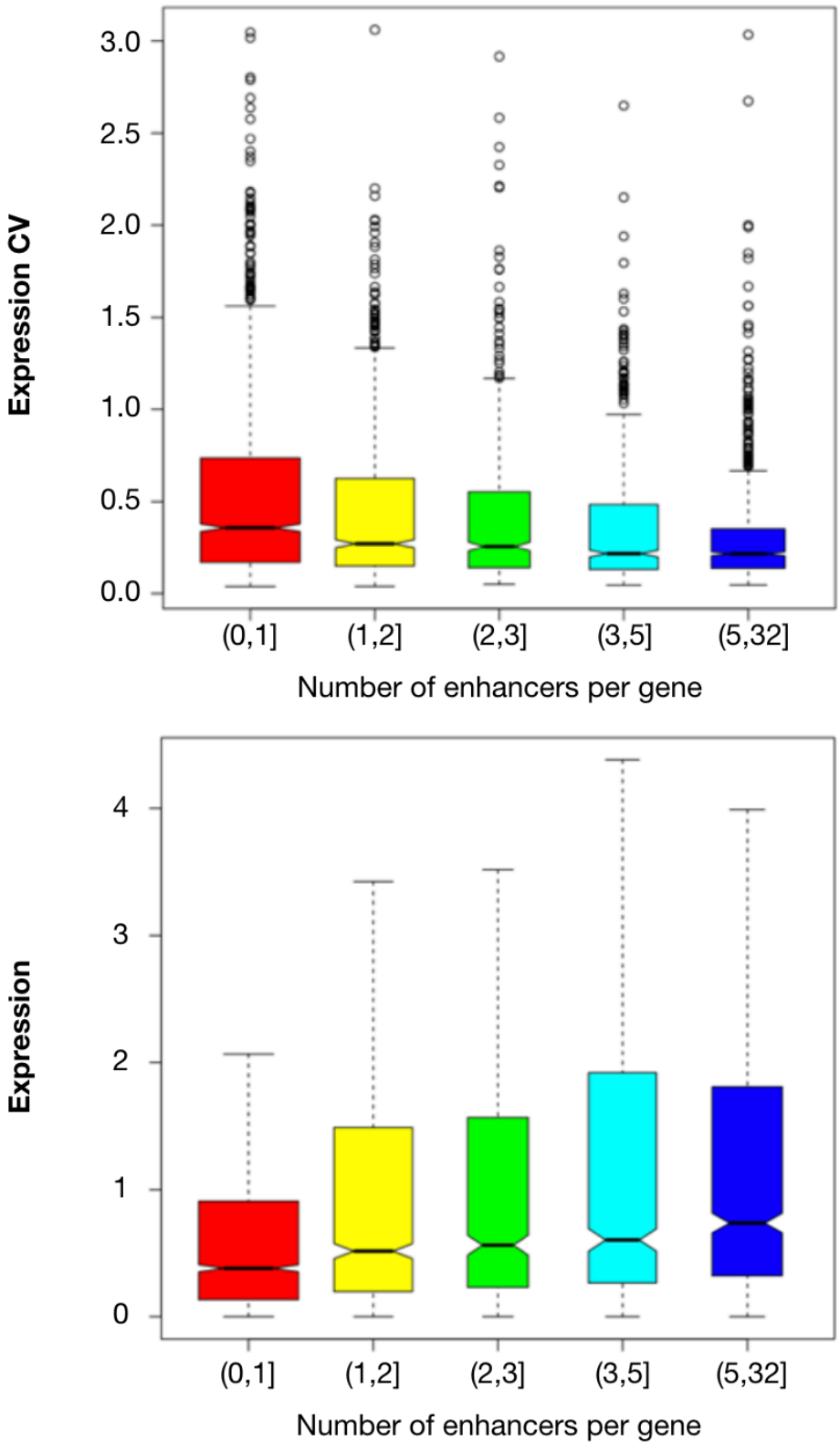


Figure 2.4: Gene expression CV (top) or expression (transcript per million) (bottom) as a function of the number of enhancers.

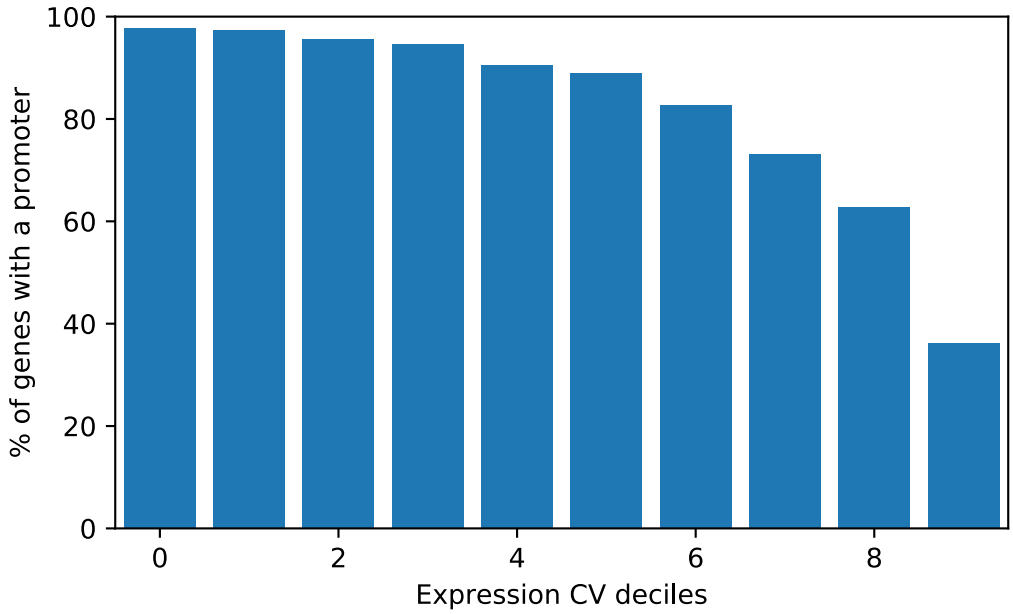


Figure 2.5: Fraction of genes with at least one promoter for top 10,000 highly expressed genes that have been grouped into deciles based on CV values (Jänes *et al* 2018).

These significant regulatory interactions provide hypotheses that can be tested. For instance, the promoter for *bec-1* is connected to an upstream enhancer and a downstream, intronic enhancer, as well as the promoter for the *skn-1c* isoform. *bec-1* and *skn-1* are induced as part of the mitochondrial unfolded protein response (UPRmt) through the binding of the bZip transcription factor ATFS-1 during mitochondrial stress (Nargund *et al* 2012, Nargund *et al* 2015). Under normal conditions, ATFS-1 is imported into mitochondria and degraded. However, under mitochondrial stress, a portion of ATFS-1 is instead imported into the nucleus to bind and induce protective genes, and down-regulate

oxidative phosphorylation and tricarboxylic acid cycle genes. This binding is contingent on the UPRmt element (UPRmtE): a 14-bp consensus element that is required for ATFS-1 binding and function (Nargund *et al* 2015). Interestingly, while the *bec-1* promoter contains two such elements, the *skn-1c* promoter does not. Whilst not definitive, the physical proximity implied from ARC-C between *bec-1* and *skn-1c* in normal conditions suggests a way by which ATFS-1 can mediate the expression of co-regulated genes involved in UPRmt despite the lack of a binding motif in one of them.

Interaction hubs

We observed that some regulatory elements participate in a large number of interactions, suggesting that they may be organisational hubs (examples in **Fig 2.8**). To investigate this, we defined hubs as elements having the top 5% of unique interaction partners (12 or more, $n = 879$). Hubs have high chromatin accessibility (**Fig 2.6**) and frequently overlap HOT regions (95.3%). 62.6% of hubs overlap promoters for protein-coding genes and, as a corollary, hubs are also often found in active chromatin state domains (62.7%, $p < 0.001$) as defined in Evans *et al* (2016). Hubs appear to be regions of high chromatin activity reminiscent of transcription factories.

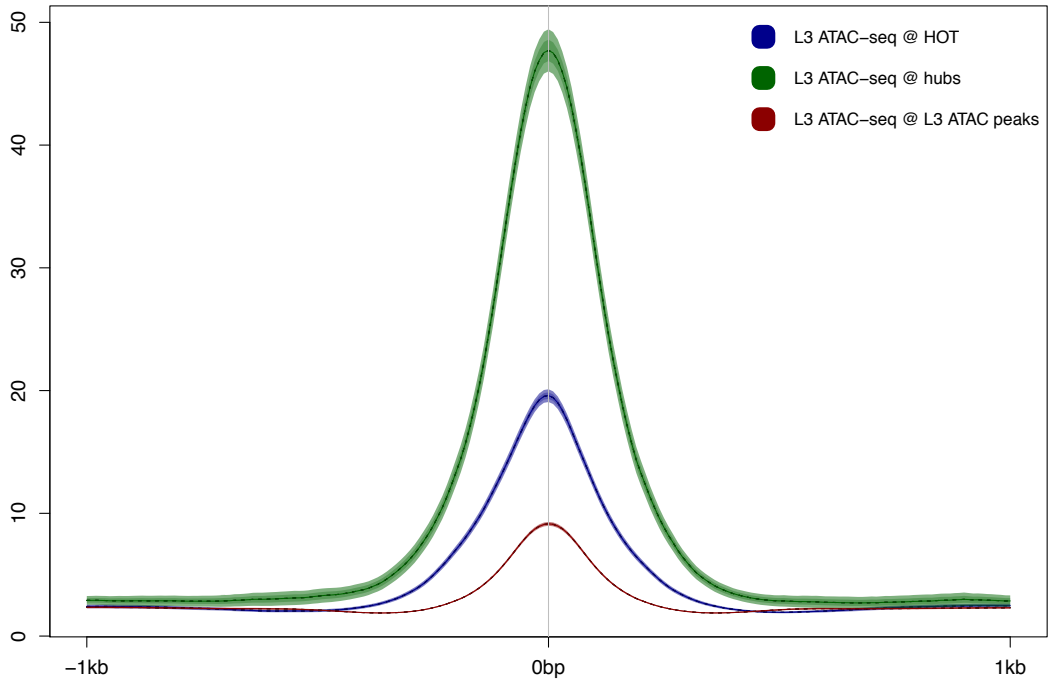


Figure 2.6: Aggregate coverage of L3 ATAC-seq centred on HOT (blue), hubs (green), or L3 ATAC-seq peaks (red).

I hypothesised that these hubs are functionally involved in the regulation of genes that they are interacting with. To test this hypothesis, I selected four strains that deleted a hub, and generated three hub deletion strains using CRISPR-Cas9 genome editing (**Method**) (**Table 2.7**; **Fig 2.8**) and tested the effect of hub deletions on local or linked gene expression. Wild-type and hub deletion strains were collected in duplicate at the L3 stage, and deletion strains were matched to wild-type by staging collections using germ line size (**Methods**). Several wild-type worm collections were made to match each mutant strain (**Table 2.7**). Genes with oscillating expression (3,845) (Hendriks *et al* 2014) were excluded in our

analyses as they are exquisitely sensitive to stage variations. Principal Component Analysis (PCA) of RNA-seq data indicated that mutant and wild-type worm collections were paired appropriately as they cluster well together (**Table 2.7**; **Fig 2.9**).

Strain	Genotype	# interacting genes	Matching wild-type stage
MT13954	*mir-81 & mir-82* (nDf54) *X.*	18	N2_E
MT16494	*mir-229 & mir-64 & mir-65 & mir-66* (nDf63) *III.*	14	N2_E
MT17429	nDf67 *IV.*	15	N2_D
ST36	*plx-1 (nc36) IV.*	14	N2_B
JA1802 (hub02)	*chd-7* (*we27*) *I.*	49	N2_E
JA1808 (hub03)	*bath-43* (*we26*) *III.*	45	N2_E
JA1800 (hub05)	K04B12.2 (*we25*) *II.*	34	N2_D

Table 2.7: Summary of hub deletion RNA-seq experiments.

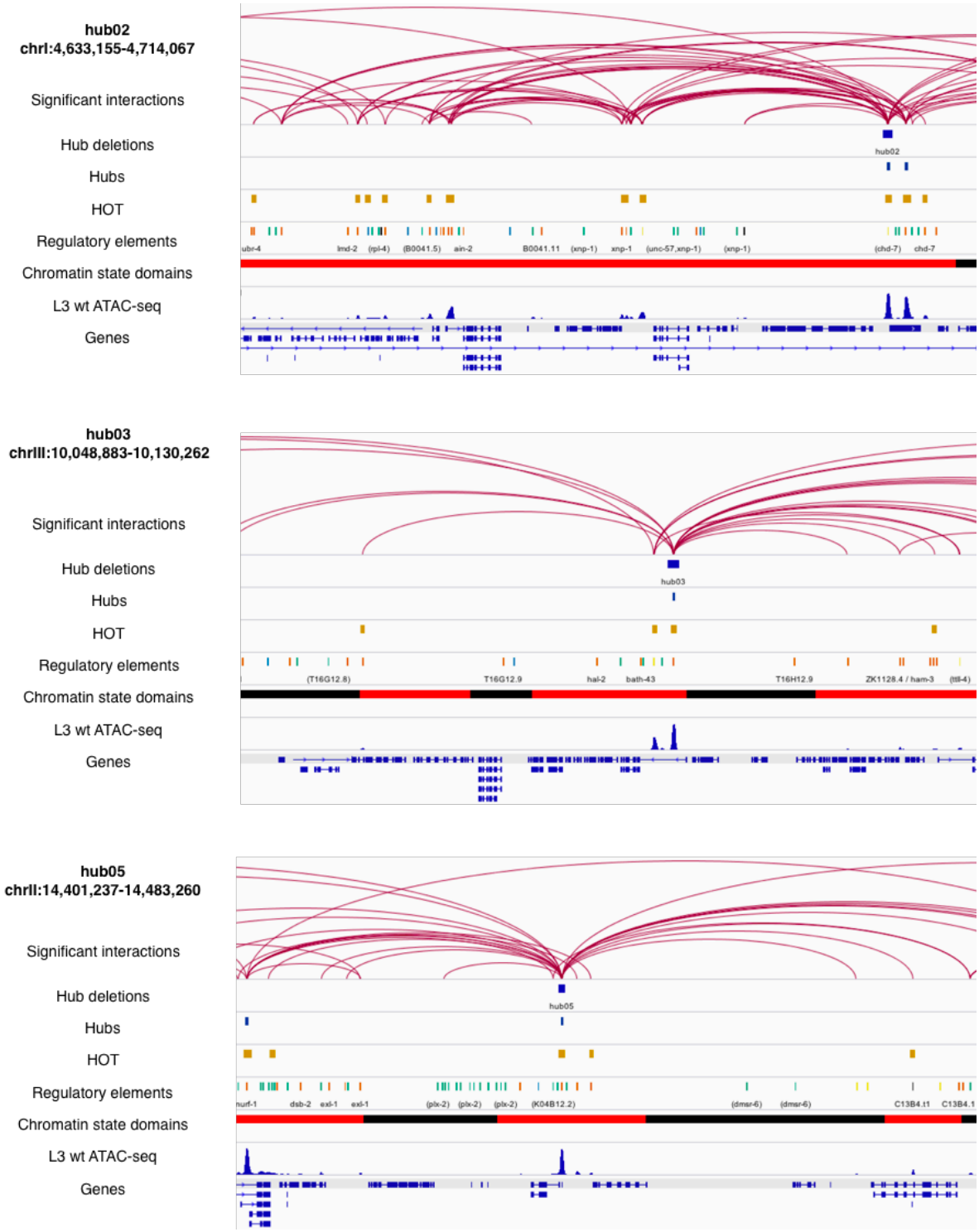


Figure 2.8: Snapshots of CRISPR-Cas9 hub deletions - hub02, hub03, hub05. Tracks show (top to bottom) significant interactions, hub deletions, hubs, HOT regions, annotated regulatory elements, chromatin state domains - active (red), regulated (black), wild-type L3 stage ATAC-seq, and genes. Hubs appear to organise clusters of significant interactions.

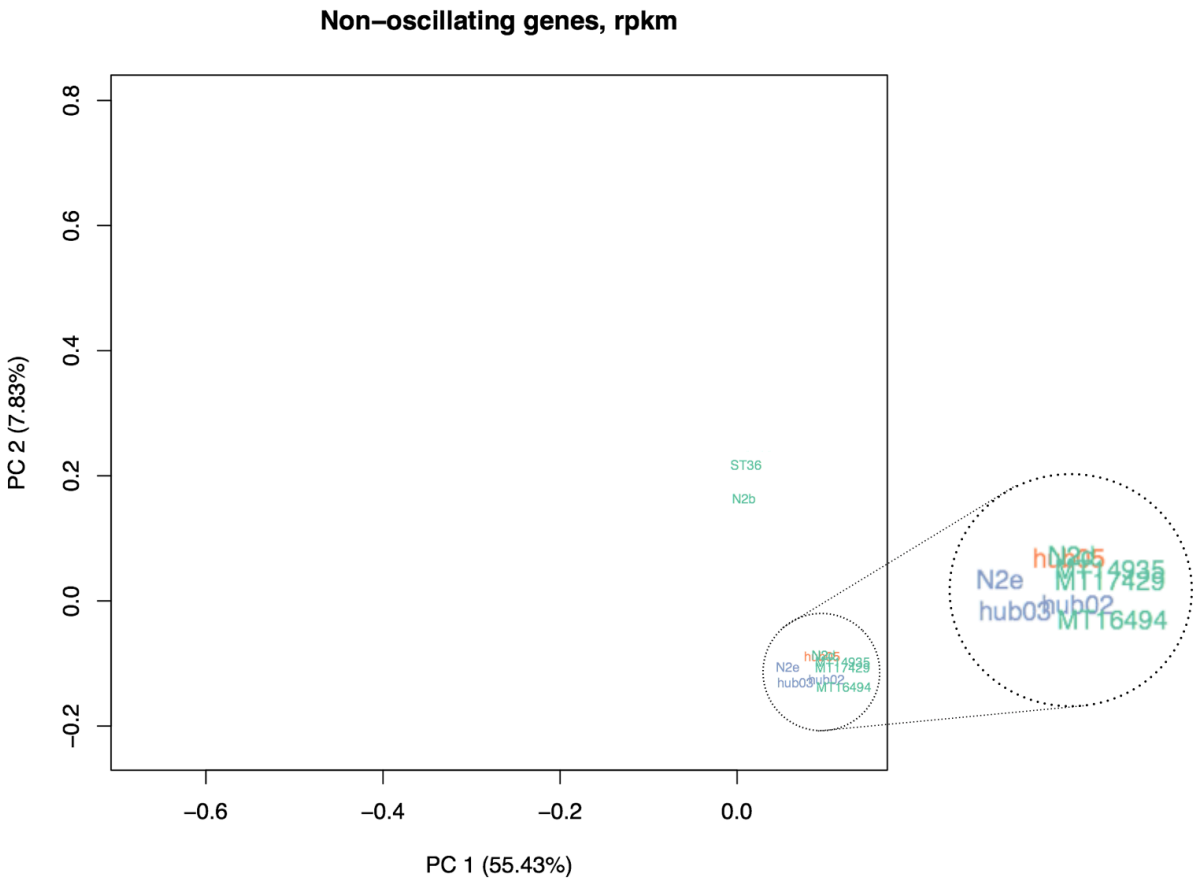


Figure 2.9: Principal component analysis of wild-type and hub deletion RNA-seq (rpkm).

Transcriptional effects of hub deletions

To evaluate the effect of hub deletions on gene regulation, we took two approaches. We determined the fold-change in expression of genes connected to the hubs ("linked genes analysis") and the fold-change in expression of genes within certain distances to the hub ("local genes analysis"). As a control, we tested if any changes were significantly different from changes seen in other regions in the genome, reasoning that if hubs regulate local or linked genes, their absence should affect these genes more than genes in other regions in the genome. For the

linked genes analysis, we compared genes linked to the hub of interest with randomly selected genes or genes linked to hubs in other deletion strains (**Appendix - Fig A1.5**) (i.e. if analysing hub02, we looked at genes connected to hub03, hub05, MT13954, MT16494, MT17429, and ST36). For the local genes analysis, we compared the variance of expression changes in windows centred on the hub deletions against randomly selected windows of the same sizes (10kb, 50kb, 100kb, 200kb, and 1Mb) or against windows used to analyse other hub deletion strains (**Appendix - Fig A1.6**). We found that none of the hub deletions specifically affected the expression of linked or local genes (one-sided t-test) as compared genes in other regions of the genome (**Appendix - Table A1.3 & Table A1.4**).

As an illustration, when we look at hub02-linked genes in hub02 mutants (**Fig 2.10** - top left panel, black circles), the three largest misregulated genes had absolute log2FC around 0.5 (representing around 1.414-fold change), but so did genes in other regions of the genome (e.g. MT16494-linked genes had 4 out of 7 genes with absolute log2FC above 0.5) (**Fig 2.10**). Overall, there were no statistically significant difference in expression fold-changes ($p = 0.549$, **Appendix - Table A1.3**) between hub02-linked genes and other linked genes in hub02 mutants. Moreover, if hub02 deletion had a local transcriptional effect, the expression fold-changes of genes would have decreasing variance with increasing

genomic distance from the deletion site. This was not the case (**Fig 2.11** - top left panel, red circles). When we compare hub02-local genes within different distances from the deletion with genes in other regions the genome (randomly selected windows while controlling for similar gene numbers), we find no statistical significance (e.g. $p = 0.302$ for 100kb, **Appendix - Table A1.4**).

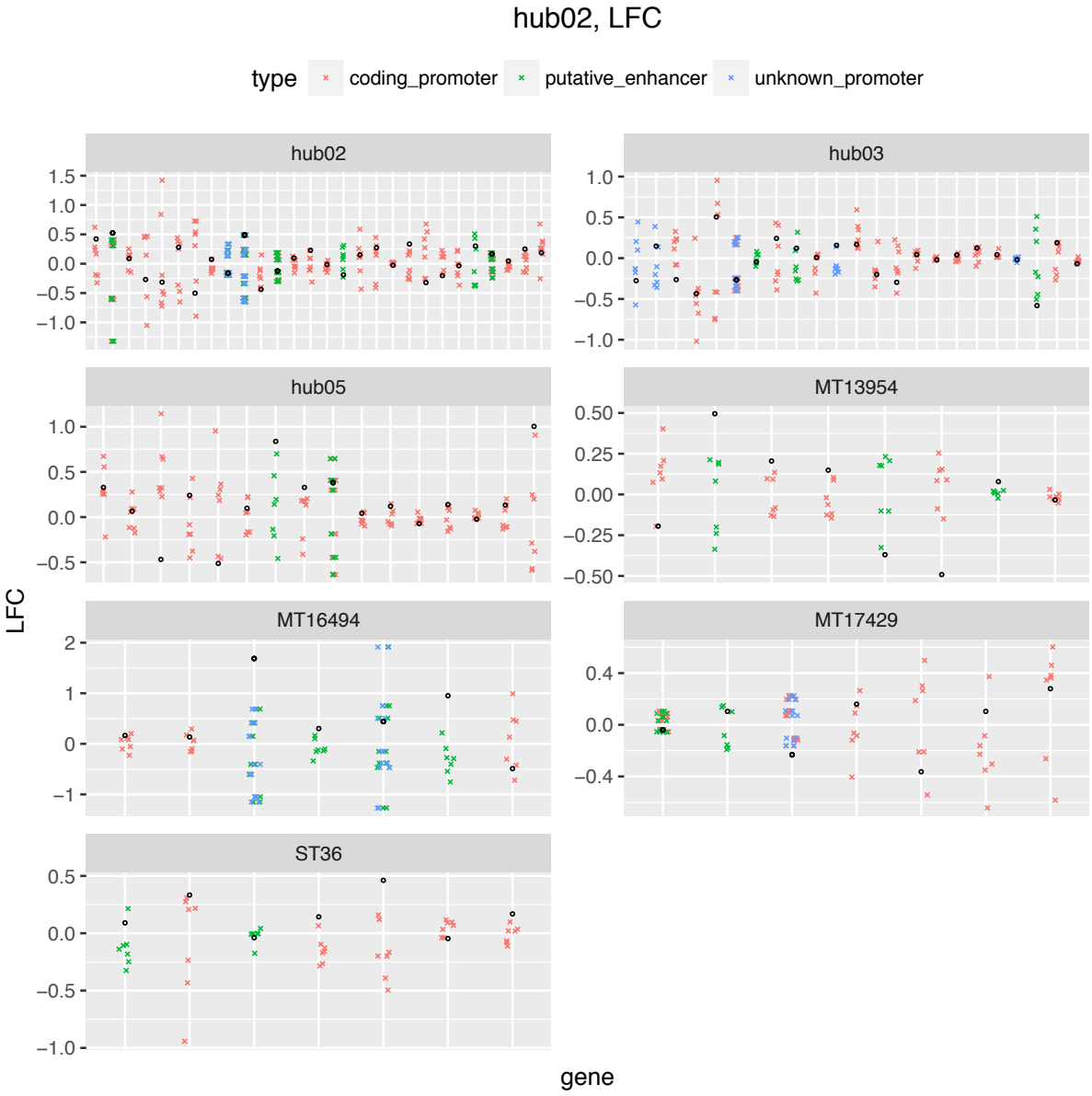


Figure 2.10: Linked-genes analysis for hub02 deletion. Expression lg2FC was calculated for each deletion's own linked genes and compared with linked genes from other strains. Black circles are lg2FCs in the deletion strain of interest, while coloured crosses are lg2FCs of the same genes in other strains.

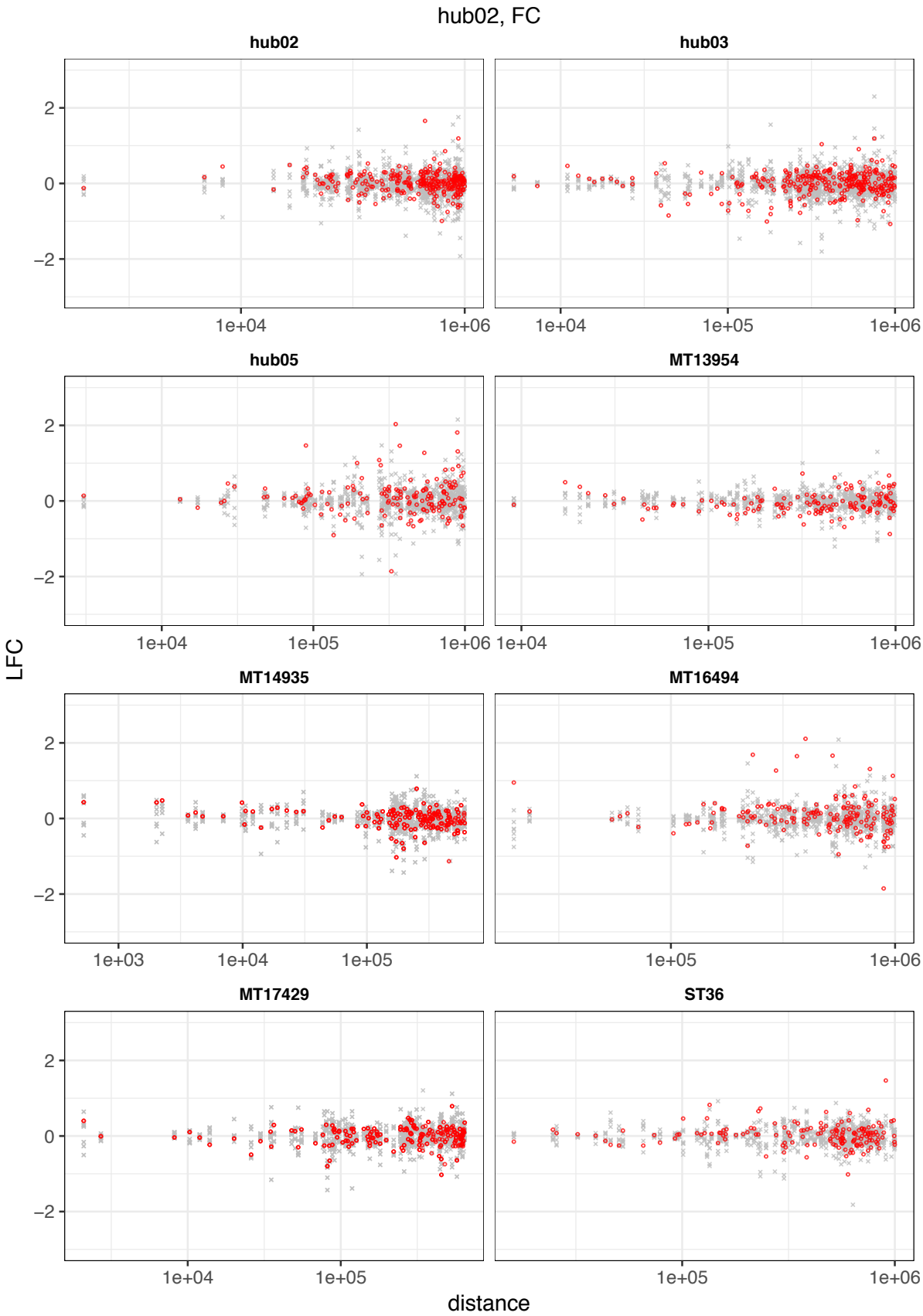


Figure 2.11: Local genes analysis for hub02 deletions. Expression lg2FC of genes is plotted as a function of genomic distance from the deletion site. Red circles

represent lg2FC of genes in the deletion strain of interest, grey crosses represent lg2FC of the same genes in other strains.

ARC-C in hub deletions

To better understand this lack of transcriptional change in the hub deletions, I decided to examine their chromatin architecture. As described previously, most hubs are HOT regions, which we found to interact indiscriminately, with pairings between promoters no better correlated than by chance (**Fig 2.3**) and co-occur with hubs, which argues for a intimate relationship between both elements. Based on this, I hypothesised that hub interactions may be redundant, such that nearby hubs or HOT regions could substitute when a hub was deleted. I chose to perform ARC-C on hub02, hub03, and hub05 deletion mutants to determine the effect of hub deletion on chromatin architecture and test this hypothesis.

Hub02 has a hub 1.2 kb away that is also a HOT region and a second close HOT region 3.2 kb (**Fig 2.8**). The next closest HOT regions are 25 kb and 27 kb away from hub02 (**Fig 2.8**). Hub03 has a HOT region 1.5 kb away, which is not defined as a hub but still in the top 15th percentile (>6) in terms of number of interacting partners (**Fig 2.8**). Finally, hub05 has a HOT region 2.6 kb away that has only one interacting partner, with the next nearest HOT regions 30 kb and 36 kb away, and no nearby hubs (**Fig 2.8**). I performed ARC-C on L3 stage larvae from hub02, hub03, and hub05. I sequenced them and pooled biological replicates

to a depth of around 5.8 million, 5.5 million, and 5.3 million informative reads respectively (**Appendix - Table A1.1**).

We used two assays to determine whether there were changes in local accessibility or interactions in hub deletion mutants. The standard deviation of log2 fold-change between hub and N2 *cis*, informative reads was calculated from 1 kb to 5 Mb away from the site of deletion. It measures change in interactions without accounting for directionality (up or down-regulated) and any rewiring of interaction (**Fig 2.12** - top) - a larger standard deviation signifies a greater amount of difference between the hub of interest and N2. This was also done for hub05 (hub05 vs hub05) and N2 (N2 vs N2; using the same window as in hub05) biological replicates as a control. As expected, they were well within the 95% confidence intervals near the median, indicating that there were no significant difference in *cis*, informative reads between biological replicates (**Fig 2.12** - top).

We observe a similar lack of significant difference for hub02 (hub02 vs N2) and hub03 (hub03 vs N2) from 1 kb to 5 Mb from the site of deletion (**Fig 2.13** - top). In contrast, for hub05, the standard deviation of log2 fold-change was significantly higher up to 100 kb, which suggests that chromatin accessibility and interactions were misregulated after the hub05 deletion (**Fig 2.12** - top). This phenomenon is decidedly local as it decays with distance (**Fig 2.12** - top).

As an alternative but related perspective, we looked at the proportion of *cis*, informative peaks with an absolute log2 fold-change above 1 in hubs vs N2 (**Fig 2.12** - bottom). Likewise, we observed a similar result - hub02 and hub03 did not show any higher proportion of peaks having more change as opposed to N2 (**Fig 2.13** - bottom). Hub05 had significantly more changed peaks up to 50 kb and again, the amount of change drops with increasing distance from the site of deletion (**Fig 2.12** - bottom).

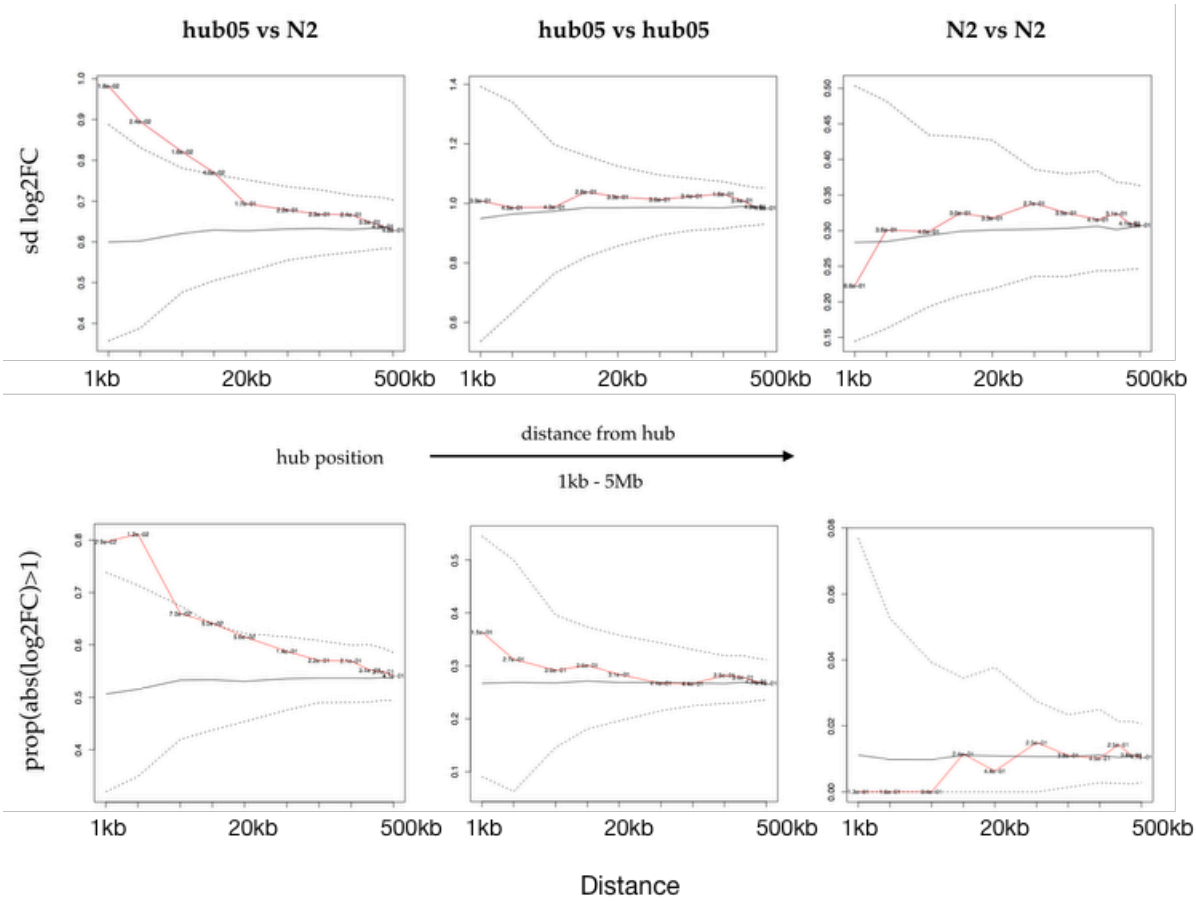


Figure 2.12: Standard deviation of log2FC (top, red line) of *cis* informative peaks between hub05 and wild-type (N2), hub05 and hub05, wild-type and wild-type. Proportion of absolute log2FC (bottom, red line) of *cis* informative peaks. Sdlog2FC and prop(abs(log2FC) were calculated for peaks at regulatory elements within 20, 40, 100, 200, 400, 1000, 2000, 5000kb of deletion. Black solid lines indicate mean sdlog2FC of 1,000

randomly sampled regulatory elements and black dashed lines indicate 95% confidence interval.

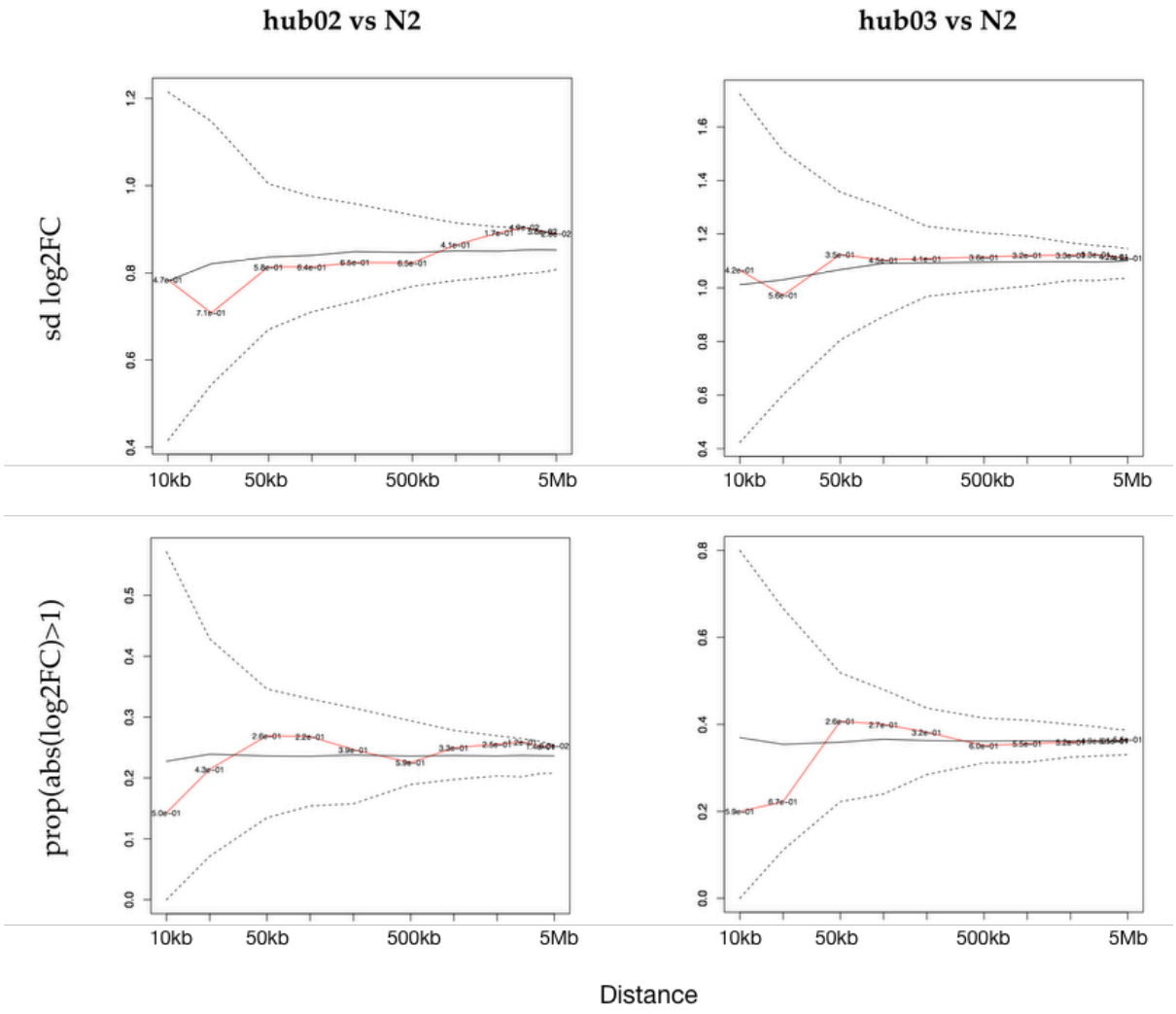
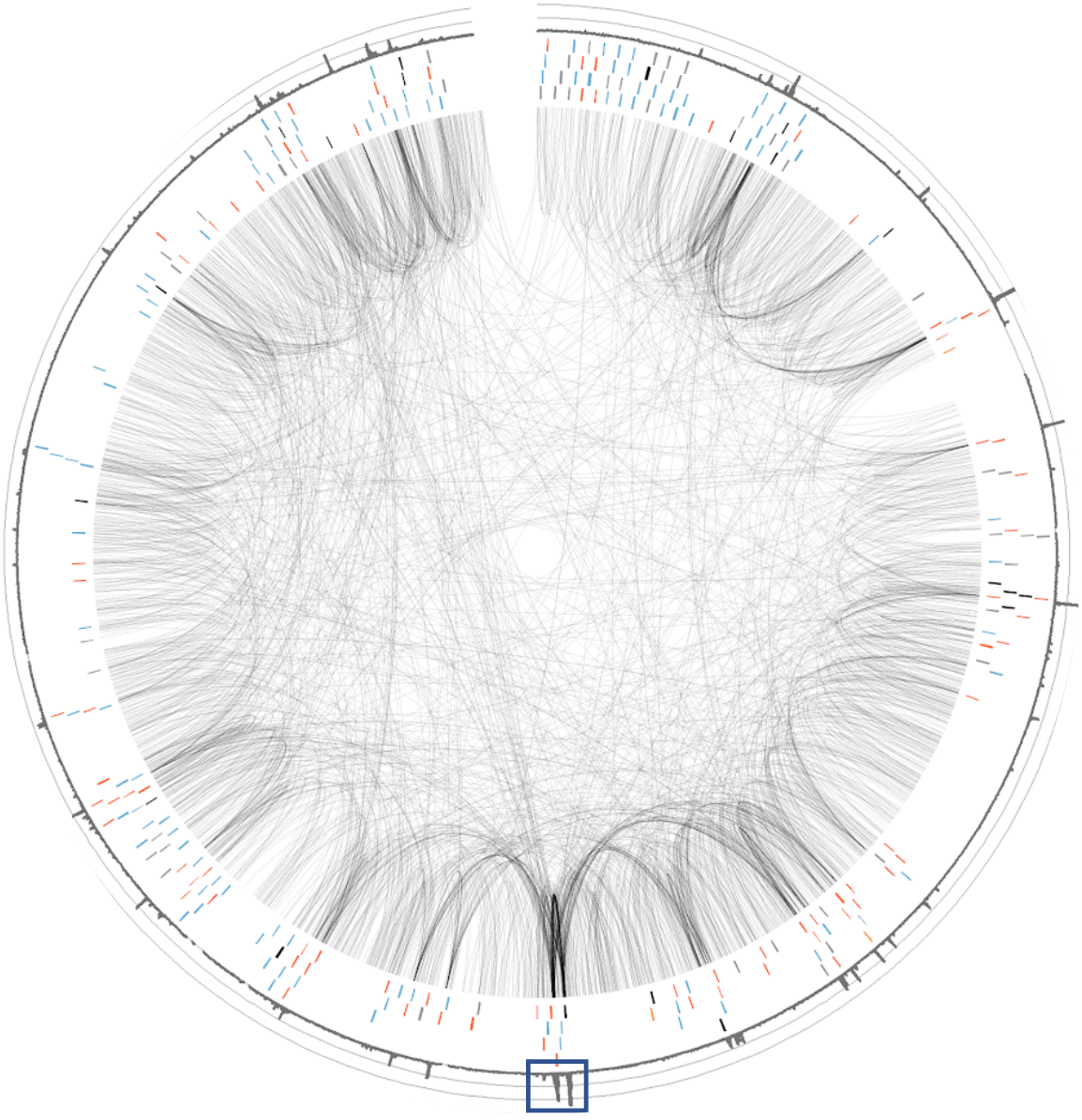


Figure 2.13: Standard deviation of log2FC (top, red line) of *cis* informative peaks between hub02 and wild-type (N2), hub03 and wild-type. Proportion of absolute log2FC (bottom, red line) of *cis* informative peaks. Sdlog2FC and prop(abs(log2FC) were calculated for peaks at regulatory elements within 20, 40, 100, 200, 400, 1000, 2000, 5000kb of deletion. Black solid lines indicate mean sdlog2FC of 1,000 randomly sampled regulatory elements and black dashed lines indicate 95% confidence interval.

Next, we examined if there were qualitative changes in local chromatin interactions as a result of hub deletions through 500 kb Circos plots of valid read pairs centred on the deletion site. We note that the deletion resulted in the loss of coverage at the expected deletion sites in all hubs (**Fig 2.14, 2.16, 2.18**: blue box encompasses coverage over deletion sites; blue line denotes region containing deleted hubs. The neighbouring hub for hub02, highly interacting node for hub03 and absence thereof in hub05, as mentioned earlier, were also captured in the Circos plots (**Fig 2.15, 2.17, 2.19**: blue box).

Qualitatively, interaction patterns seem to be preserved in both hub02 and hub03, while certain enriched interactions appear to be lost in hub05 (**Fig 2.14**, **2.16**, **2.18**), which is consistent with our analyses using *cis,* informative reads. To illustrate this more precisely, for the hub02 window, the viewpoint hub region 5 in N2 has enriched interactions with regions 1, 2, 3, 4, 6, and 7 (**Fig 2.15**). This is also reproduced in hub02, but with interactions coming from the adjacent hub (**Fig 2.15**). Strong interactions within local non-hub regions are also recapitulated, such as within regions 4 and 7 in both N2 and hub02 ARC-C (**Fig 2.15**). We observe the same phenomenon in hub03 mutants. Interactions emanating from the hub of interest in region 5 to regions 1, 2, 3, and 4 have shifted to the adjacent peak (**Fig 2.17**). In contrast, enriched interactions connecting regions 1, 2, and 4

with the viewpoint hub region 3 in N2 were lost in hub05 mutants (**Fig 2.19**). The local chromatin interaction landscape in hub02 and hub03 appears to be largely unchanged with interactions shifting to an adjacent hub or highly interacting region, whereas the loss of hub05 appears to have precipitated a loss of strong interactions.



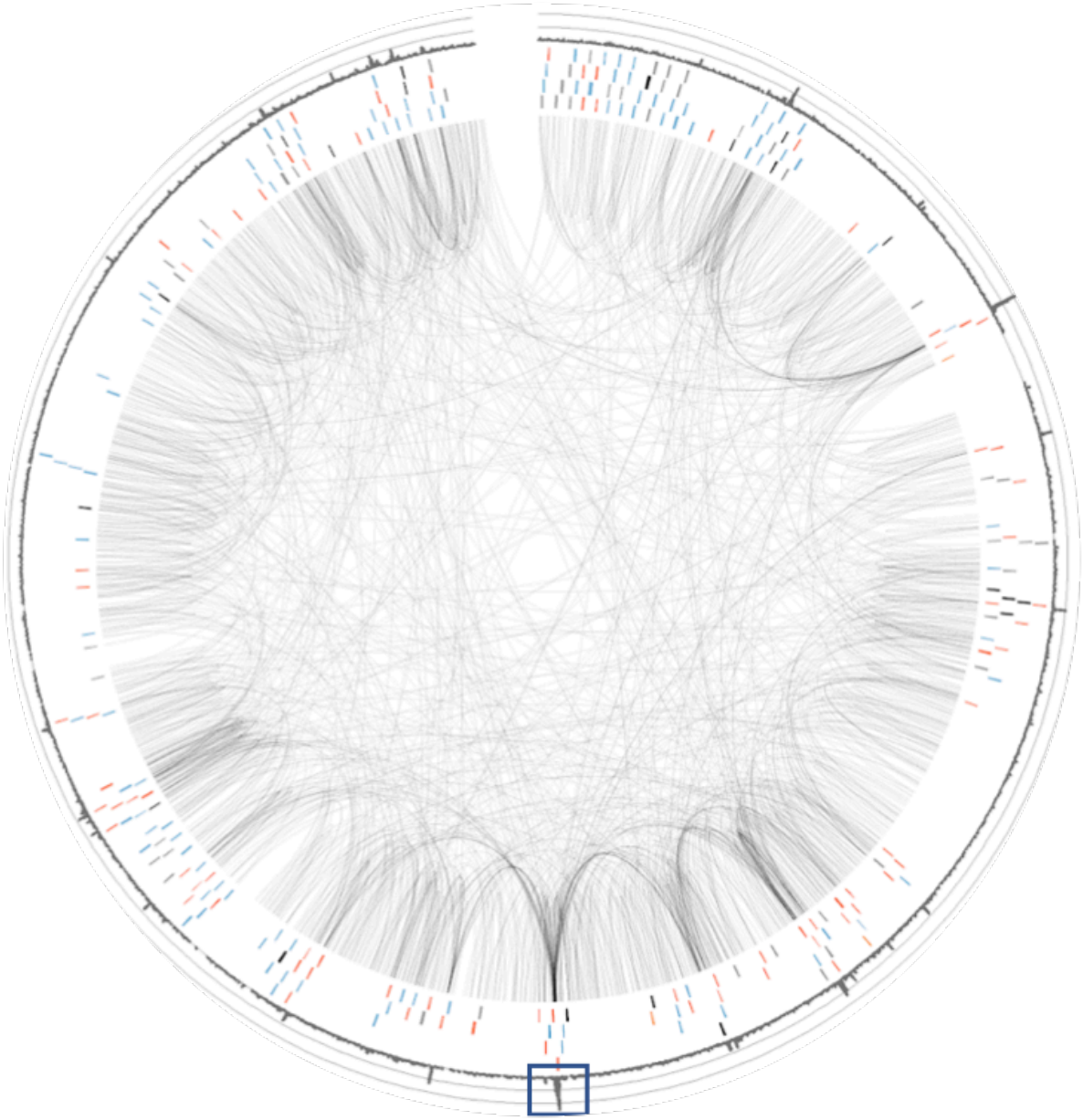


Fig 2.14: **Wild-type** (top) and **hub02** (bottom). Outer to inner ring: 2D informative read coverage, annotated regulatory elements, informative interactions. Blue box and line indicate region containing hub deletion. Chr I: 4,450,000-4,950,000.

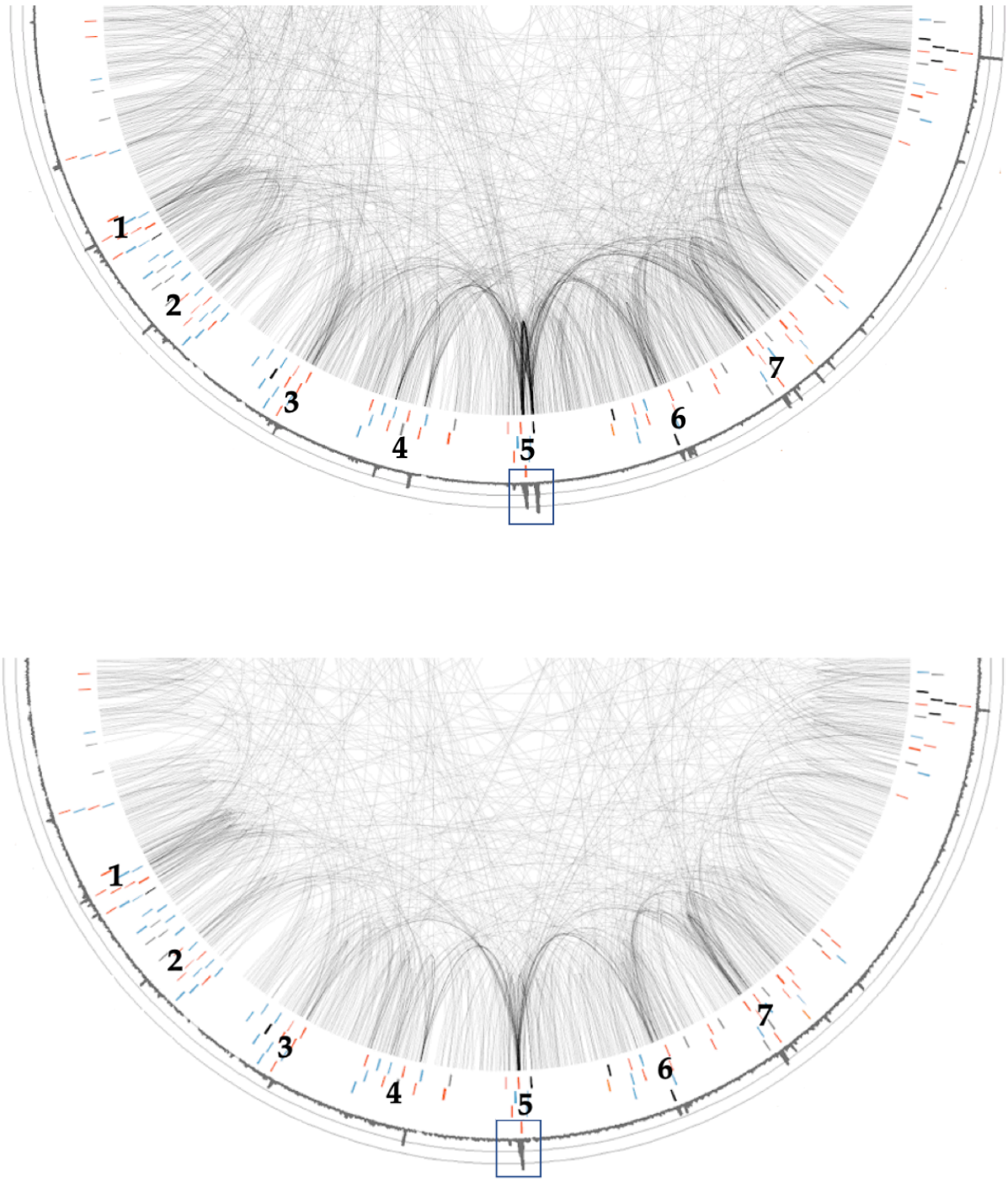
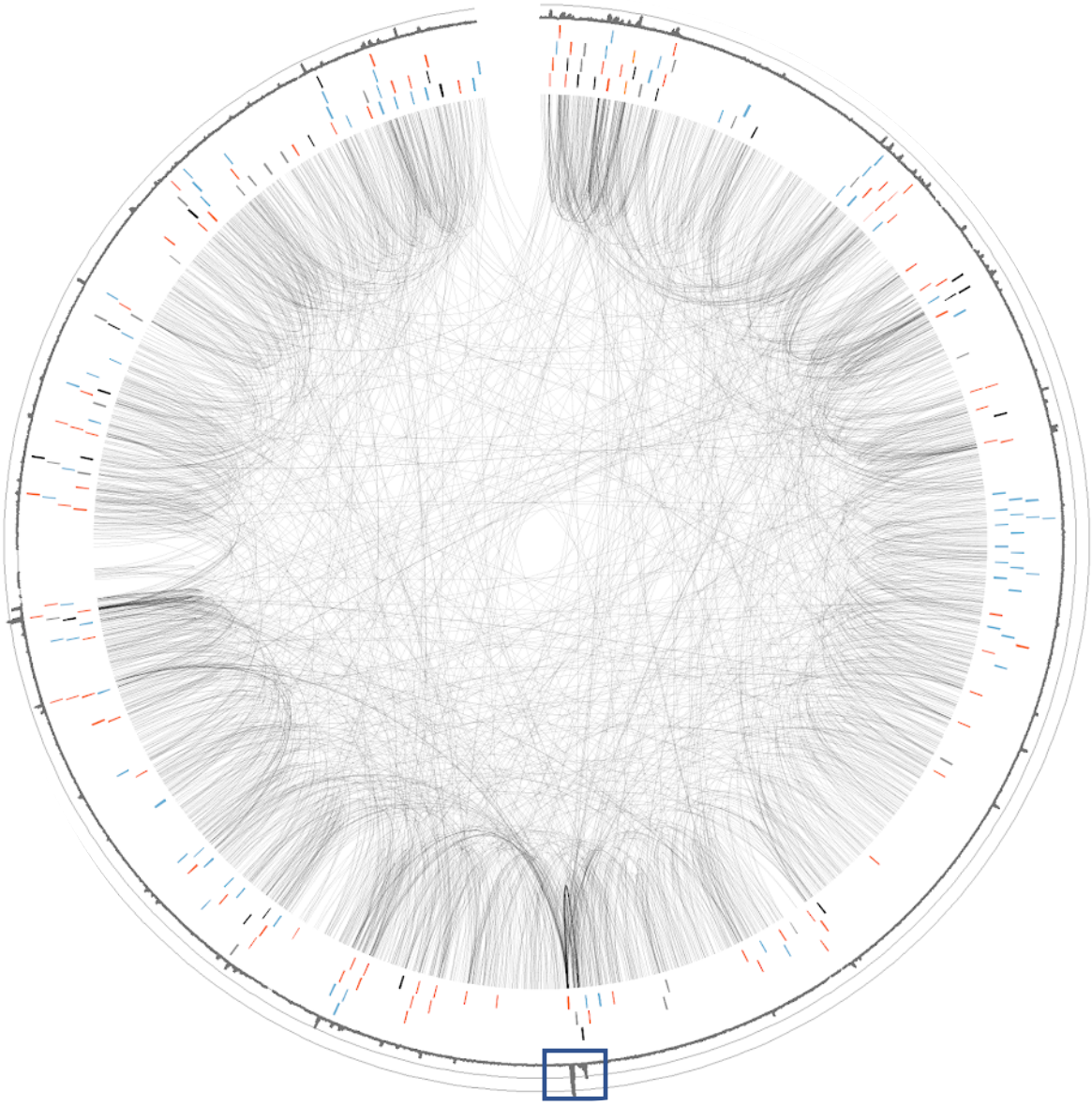


Fig 2.15: **Wild-type** (top) and **hub02** (bottom). Outer to inner ring: 2D informative read coverage, annotated regulatory elements, informative interactions. Blue box and line indicate region containing hub deletion. Chr I: 4,575,000-4,825,000.



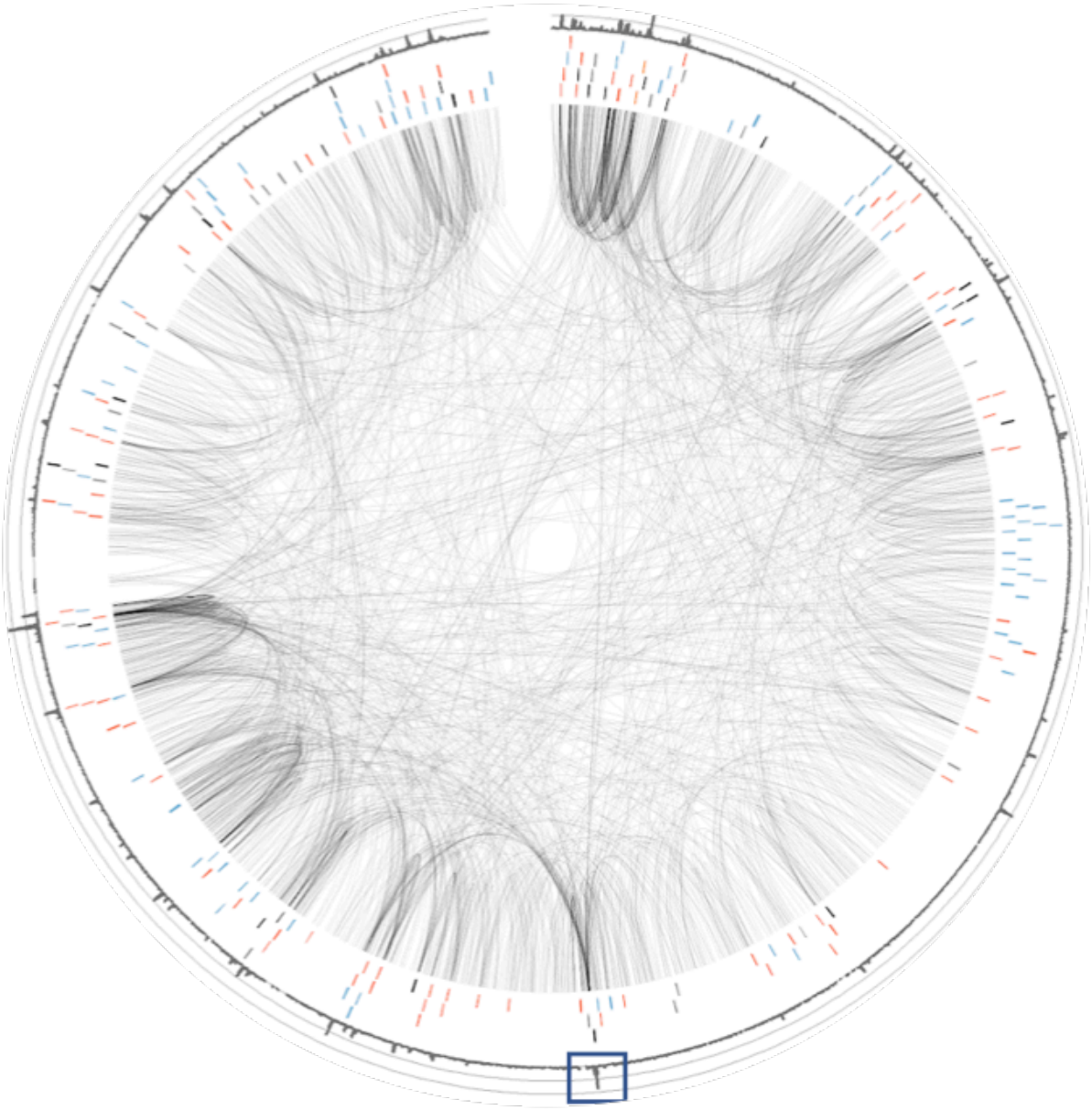


Fig 2.16: **Wild-type** (top) and **hub03** (bottom). Outer to inner ring: 2D informative read coverage, annotated regulatory elements, informative interactions. Blue box and line indicate region containing hub deletion. Chr III: 9,850,000-10,350,000.

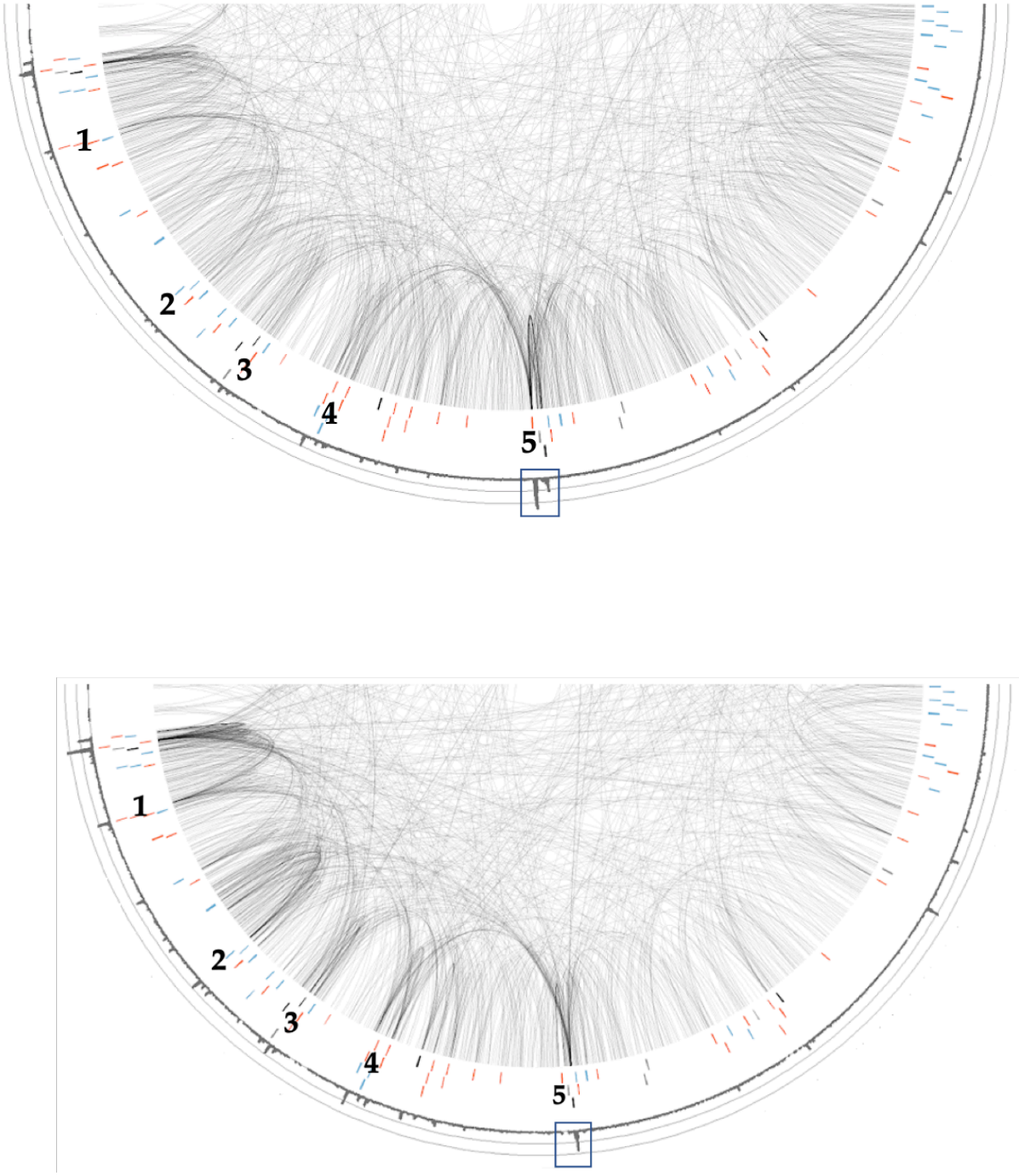
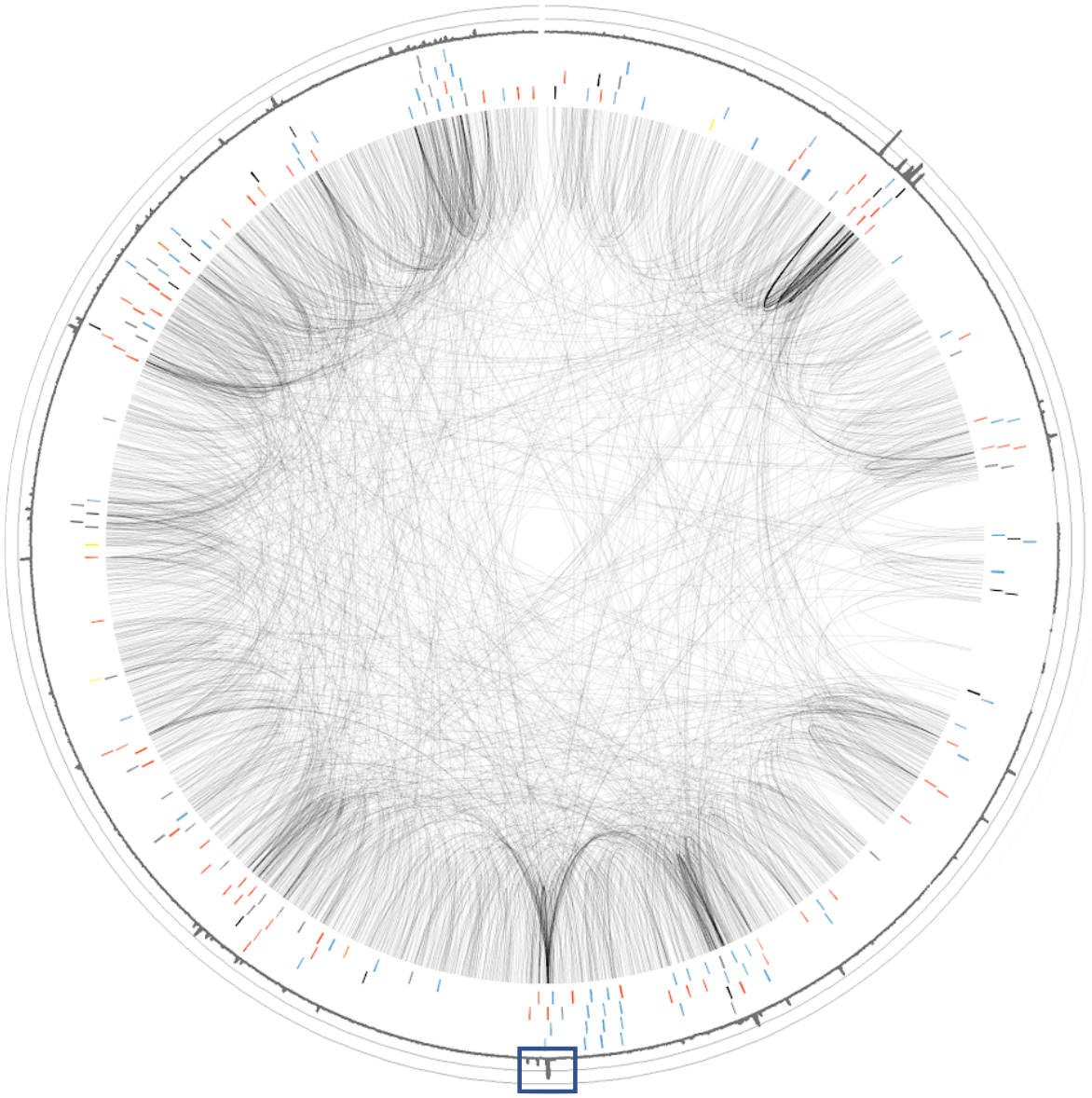


Fig 2.17: **Wild-type** (top) and **hub03** (bottom). Outer to inner ring: 2D informative read coverage, annotated regulatory elements, informative interactions. Blue box and line indicate region containing hub deletion. Chr III: 9,975,000-10,225,000.



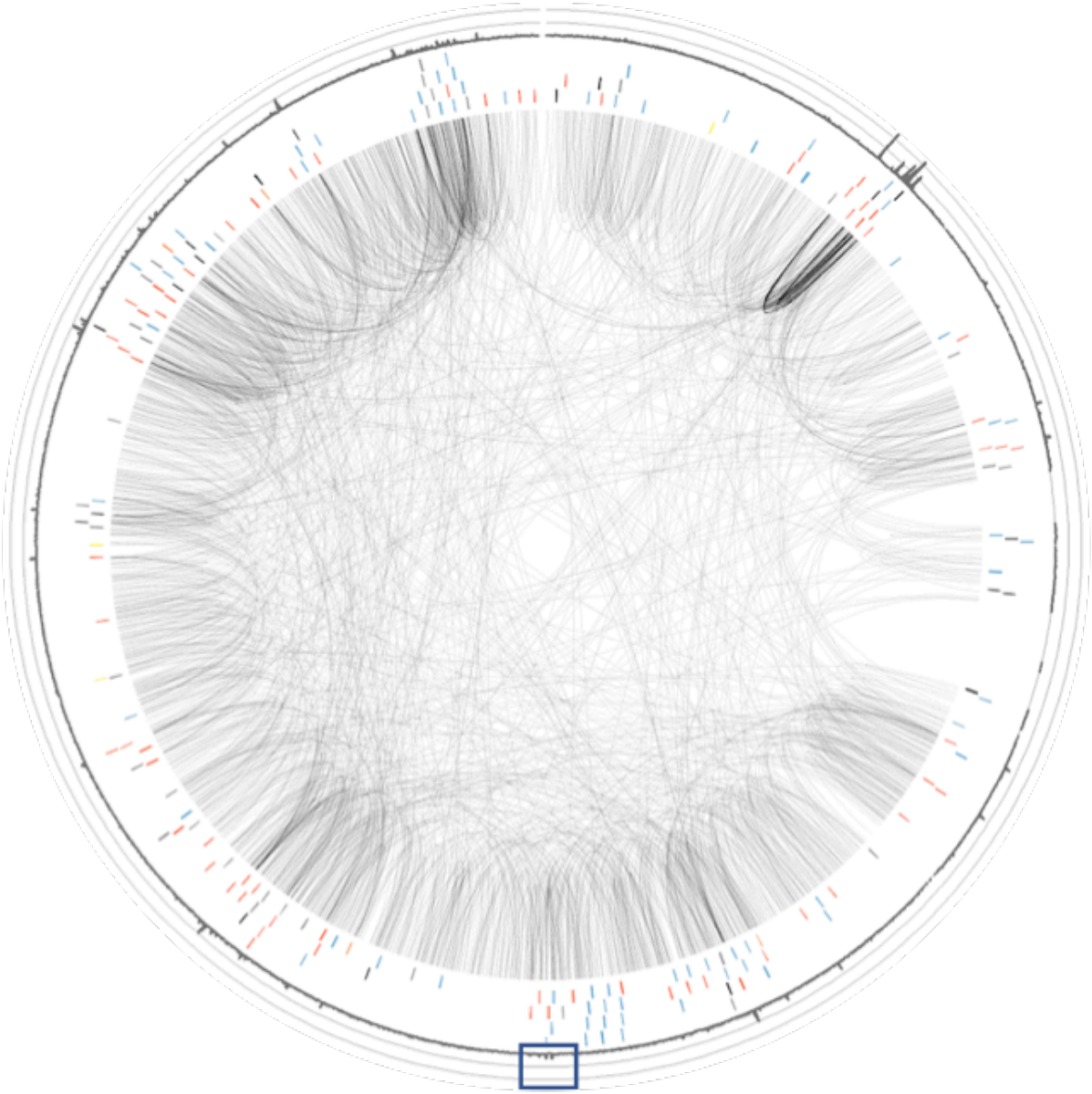


Fig 2.18: **Wild-type** (top) and **hub05** (bottom). Outer to inner ring: 2D informative read coverage, annotated regulatory elements, informative interactions. Blue box and line indicate region containing hub deletion. Chr II: 14,150,000-14,650,000.

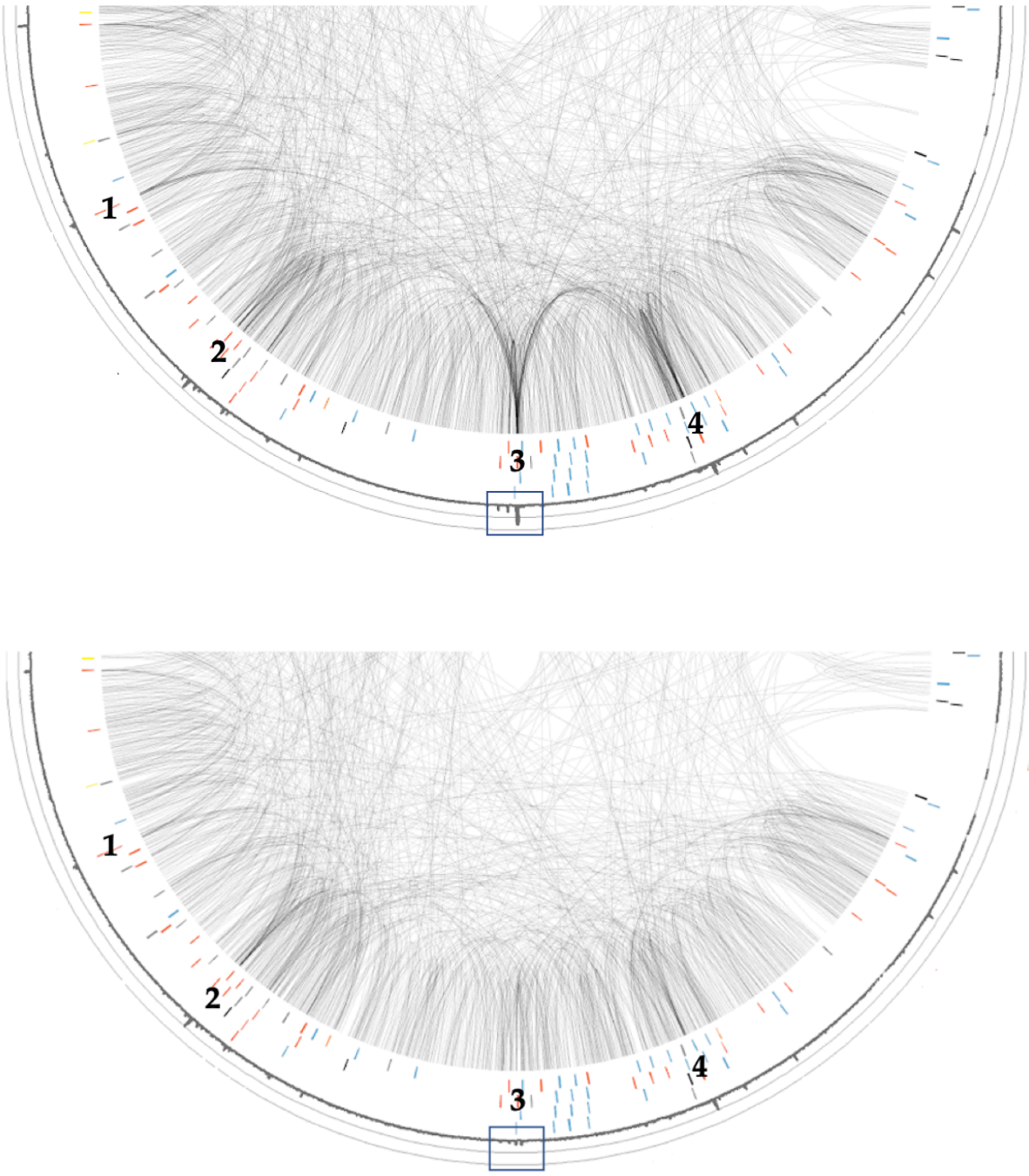


Fig 2.19: **Wild-type** (top) and **hub05** (bottom). Outer to inner ring: 2D informative read coverage, annotated regulatory elements, informative interactions. Blue box and line indicate region containing hub deletion. Chr II: 14,275,000-14,525,000.

By examining local informative coverage and the pattern of local valid interactions, we found that local chromatin interactions were changed in hub05, but not in hub02 and hub03 mutants. This is likely due to redundancy, where nearby hub or HOT regions substitute for the deleted hub. While there was a nearby HOT region for hub05, it was likely not active, as evinced by the lack of ATAC-seq peak (**Fig 2.8**). It is worth noting that HOT regions are defined from ChIP-seq in multiple developmental stages and thus reflect an organism but not stage-specific feature. We are currently working on a quantitative approach to analyse the rewiring of chromatin interactions.

It is surprising that the change in chromatin interactions in hub05 mutants do not affect gene expression. There is, however, precedent - the depletion of cohesin via auxin-inducible degron system resulted in extensive loss of loops and looped domains but had only modest effects on transcription (Rao *et al* 2017), suggesting that hubs could be structural. The effect of hub deletions might only be apparent under certain conditions, such as stress. Moreover, the loss of one hub does not imply the complete absence of access of linked or local genes to transcription machinery. It might require multiple hub or HOT deletions before any transcriptional effect is apparent.

CHAPTER III: FACTORS MEDIATING LOOP FORMATION

Cohesin and CTCF are the best studied proteins in the field of chromatin architecture for their role in looping. Beyond that, very little is known about the other factors mediating loops, domains, and compartments (as reviewed in **Introduction**). Cohesin and CTCF are often found at both ends of loops, particularly at TAD corner peak foci (Rao *et al* 2014). By extension, other factors that mediate loops are also expected to have the same property. Therefore, we screened for factors that are enriched for being at both ends of loops, reasoning that they would have a higher potential to be involved in looping in some form or manner.

Factor APA

To do so, we performed APA using paired permutations of ChIP-seq peaks within 20kb-1.5Mb, using both ARC-C and Hi-C data at 1 kb resolution. To obtain candidates, we curated a high-quality collection of 83 ChIP-seq datasets of transcription factors, chromatin regulators, and structural proteins from modENCODE, modERN, our in-house databases, and other sources (Dernburg -

modENCODE, Kranz *et al* 2013, Wiesenfahrt *et al* 2016, Latorre *et al* 2015, Kudron *et al* 2018). We used modENCODE and modERN peak calls and those from publications and called peaks on our unpublished datasets; we only used datasets with at least 300 peak calls.

GFP-tagged ChIP-seq datasets frequently do not overlap very well with endogenous protein ChIP-seq. Anti-GFP antibodies may have artefactual binding and a bias toward HOT regions; peaks were found at HOT regions that were not otherwise present in protein ChIP-seq (Kudron *et al* 2018). To remove the potential source of bias from GFP-tagged datasets and to obtain a conservative, rigorous set of binding peaks, we defined a set of 'cold' peaks - fewer than 7 factors bindings - and kept only cold peaks in every GFP-tagged ChIP-seq dataset (**Table 3.2**: "modENCODE-cold", "modERN-cold"). For factors with endogenous protein ChIP-seq, the criteria for passing our screen are to have an APA fold-change (FC) score 1.1-fold higher than that of randomly paired regulatory elements and a false discovery rate corrected p-value below 0.05. modENCODE and modERN "cold" ChIP-seq datasets were compared to randomly paired "cold" ATAC-seq peaks.

34 factors out of 78 tested were enriched for being at both ends of interactions, comprising subunits of cohesin, condensin, chromatin regulators,

and transcription factors (summarised in **Table 3.2**). In all instances, ARC-C outperforms Hi-C in relation to the number of factors that passed our criteria (**Fig 3.1**) and APA FC scores which shows ARC-C's ability to identify interactions between regulatory elements and increased sensitivity for this assay (**Fig 3.1**; examples of individual APA plots in ARC-C and Hi-C are shown in **Fig 3.3**). As expected, *rex-rex* interactions were the strongest in both ARC-C and Hi-C (**Fig 3.1**). Factors that were identified in Hi-C, with the exception of CEH-28, were also found in ARC-C (**Fig 3.1**). Interestingly, many of these were subunits from the three condensin complexes (SMC-4, HCP-6, KLE-2, MIX-1) (**Table 3.2**), providing strong evidence that condensin mediates loops in *C. elegans*.

The removal of HOT regions from modENCODE and modERN datasets is likely to have caused some factors to fail to show significant enrichment at interaction ends because some real binding sites will also have been removed. As examples, APA scores from cold peaks for ELT-2 and LIN-35 are 2.100 and 1.377 respectively at 20kb-1Mb, while scores from using all peaks from untagged ELT-2 and LIN-35 ChIP-seq produced 3.656 and 2.209 respectively (**Table 3.2**).


Crane Hi-C
20kb-1.5Mb dist
1kb resolution
ARC-C
20kb-1.5Mb dist
1kb resolution
random_TFBS.MODENCODE_hot
random_TFBS.MODENCODE_noncold
random_ATAC.Jurgen
random_TFBS.MODENCODE_cold
rex.Crane15
rec.Albritton17
ELT-2_L3.Wiesenfahrt16
CAPG-2_EM.Kranz13
SMC-4_EM.Kranz13
HCP-6_EM.Kranz13
PQN-85_EM.Kranz13
KLE-2_EM.Kranz13
JMJD-2_L3.JADB
SCC-1_L3.JADB
ZFP-1_L3.JADB
MDT-31_L3.JADB
GFL-1_L3.JADB
EPC-1_L3.JADB
RNT-1_L1.modERN_cold
PQM-1_L3.modENCODE_cold
KLE-2_L3.Kranz13
FOS-1_L2.modENCODE_cold
ELT-3_L1.modENCODE_cold
DOT-1.1_L1.JADB
MIX-1_EM.Kranz13
NHR-23_L3.modENCODE_cold
LIN-37_L3.JADB
HLH-1_EM.modENCODE_cold
NURF-1_L3.JADB
LIN-53_L3.JADB
SIN-3_L3.JADB
ELT-2_L3.modERN_cold
LIN-40_YA.JADB
LIN-15B_L1.JADB
HCF-1_L3.JADB
NHR-80_YA.modERN_cold
DAF-16_L4.modENCODE_cold
LIN-35_L3.JADB
BLMP-1_L1.modENCODE_cold

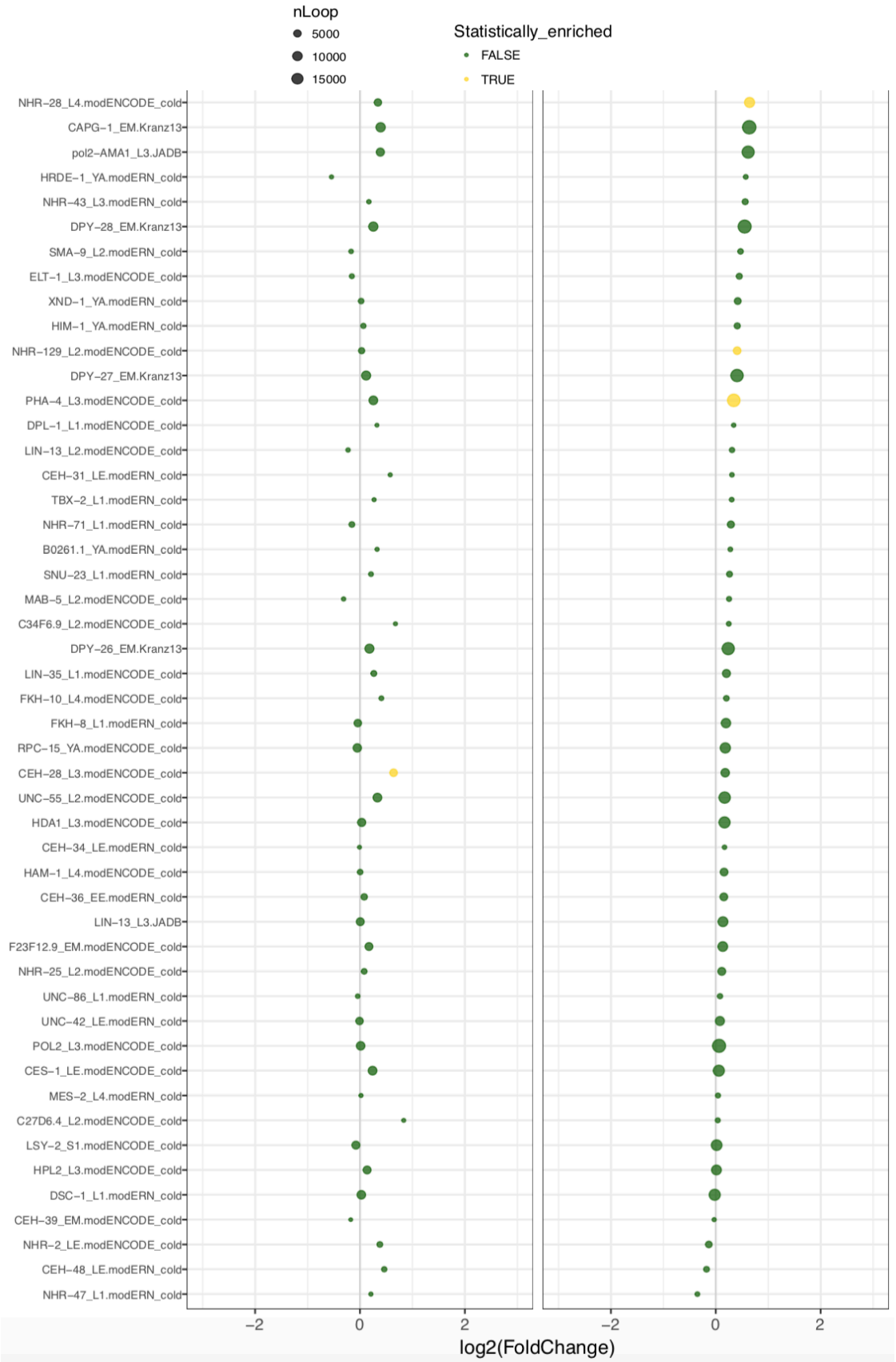


Figure 3.1: Summary of factor APA in Crane Hi-C (left) and ARC-C (right). Factors that have an APA score 1.1-fold higher than randomly paired ATAC peaks (third row) and FDR corrected $p < 0.05$ are shown in yellow.

Factor	Function	Category	Source	APA ARC-C
SCC-1	Rad21/Rec8-like family of cohesion proteins	cohesin	JA DB	1.928
COH-1	Rad21 homolog	cohesin	Dernburg	1.568
PQN-85	Nipbl ortholog; cohesin loading factor	cohesin	Kranz	2.840
MIX-1	Smc2 homolog; chromosome segregation, dosage compensaion	condensin	Kranz	2.109
SMC-4	Smc4 homolog; interacts with MIX-1 for chromosome segregation	condensin	Kranz	3.018
KLE-2	Ncaph2 homolog; mitotic sister chromatid segregation	condensin	Kranz	2.771
CAPG-2	Ncapg2 homolog; mitotic sister chromatid segregation	condensin	Kranz	3.142
HCP-6	chromosome condensation, sis chromatid segregation, microtubule attachment	condensin	Kranz	2.897
MDT-31	Med31 ortholog; RNA pol II cofactor activity	mediator complex	JA DB	2.044
ZFP-1	Af10 homolog	ZFP-1/DOT-1.1 (AF10/DOT1)	JA DB	2.746
GFL-1	Gas41 ortholog; similar to Af9 and Enl	SWR1/SRCAP	JA DB	2.523
EPC-1	Epc1/2 (enhancer of polycomb), PcG family; affinity purified with DP1, E2F6, EZH2, Sin3B in mice	TIP60/NuA4 E2F6 SIN3?	JA DB	2.414
DOT-1.1	Dot1L ortholog; H3K79 methyltransferase	ZFP-1/DOT-1.1 (AF10/DOT1)	JA DB	2.236
NURF-1	Nurf301 ortholog	NURF	JA DB	2.168
SIN-3	Sin3 family of histone deacetylase subunit	SIN3-RPD3	JA DB	2.150

LIN-40	Mta1 homolog; vulva cell fate specification and morphogenesis	NuRD	JA DB	2.180
LIN-15B		DREAM?	JA DB	2.082
HCF-1	cell cycle regulation and mitotic histone modification; interacts with SIN3 HDAC to regulate transcription	COMPASS/SIN3?	JA DB	2.082
HLH-1	Myogenic regulatory factor (MRF) ortholog; bHLH TF	TF	modENCODE-cold	2.136
ELT-2	GATA-type TF similar to Gata4-6; gut differentiation	TF	Wiesenfahrt	3.656
ELT-2	GATA-type TF similar to Gata4-6; gut differentiation	TF	modENCODE-cold	2.100
ELT-3	GATA-type TF; hypodermal cell differentiation	TF	modENCODE-cold	2.292
LIN-35	Rb ortholog; class B synMuv gene	DREAM NuRD	Latorre	2.209
LIN-35	Rb ortholog; class B synMuv gene	DREAM NuRD	modENCODE-cold	1.377
LIN-53	Rbbp4 (RbAp48) homolog; class B synMuv gene	DREAM NuRD SIN3	Latorre	2.811
NHR-23	nuclear hormone receptor; DNA binding; larval molts	TF	modENCODE-cold	1.953
NHR-28	nuclear hormone receptor; DNA binding	TF	modENCODE-cold	1.735
NHR-80	nuclear hormone receptor; DNA binding; regulates fatty acid metabolism	TF	modERN-cold	2.019
NHR-129	nuclear hormone receptor; DNA binding	TF	modENCODE-cold	1.395
PHA-4	FoxA TF	TF	modENCODE-cold	1.434
DAF-16	FoxO homolog; insulin/IGF-1-mediated signalling pathway	TF	modENCODE-cold	1.957
BLMP-1	Blmp1 ortholog; SET domain-containing	TF	modENCODE-cold	1.678

FOS-1	bZip TF	TF	modENCODE-cold	2.152
PQM-1	C2H2-type zinc finger and leucine zipper-containing protein; stress-response	TF	modENCODE-cold	2.338
RNT-1	Runx family of transcriptional regulators; developmental processes	TF	modERN-cold	2.275

Table 3.2: Description and summary of factors that passed criteria. These factors have APA scores 1.1-fold higher than randomly paired ATAC peaks and FDR p-value < 0.05.

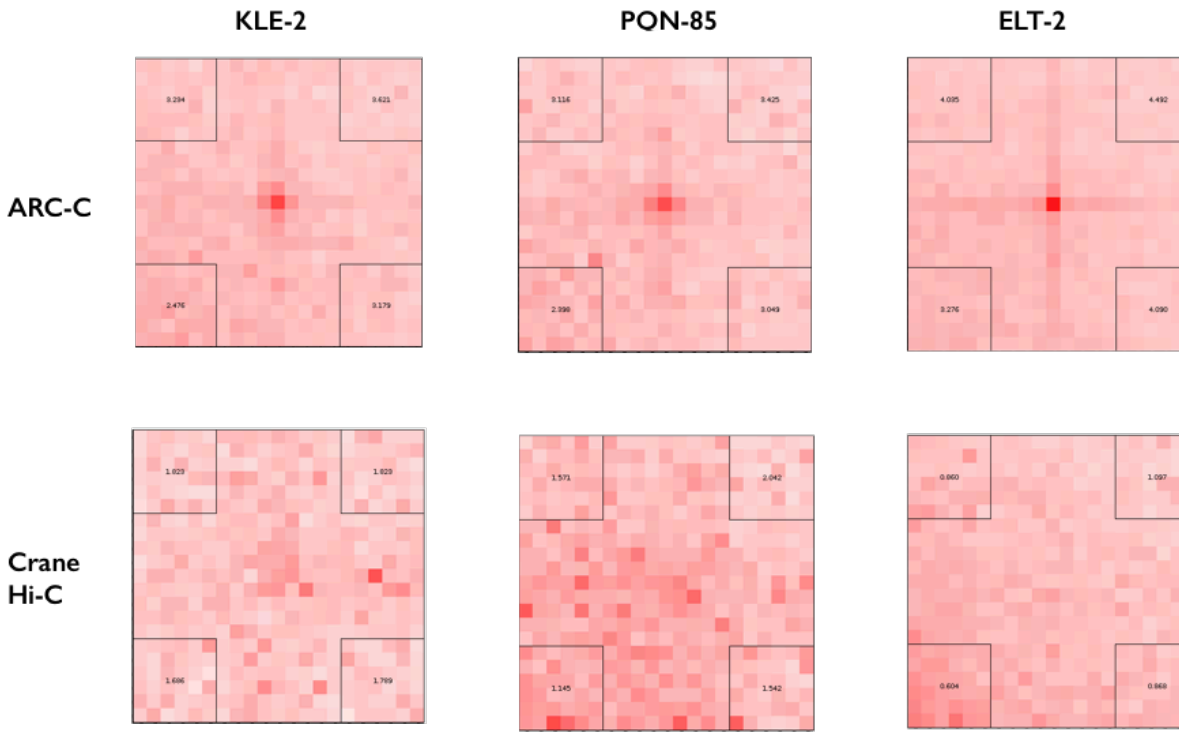


Figure 3.3: APA plots of KLE-2, PQN-85, and ELT-2 paired peaks within 20kb-1.5Mb in ARC-C (top) and Crane Hi-C (bottom).

Cohesin and condensin

Cohesin and condensin subunits constituted 6 out of 8 of the highest enriched factors in our analysis. We found subunits of the cohesin complex (COH-1, SCC-1/COH-2) and its loading factor (PQN-85, a ortholog of yeast Scc2p, *Drosophila* NIPPED-B, and human NIPBL). Cohesin is thought to be a motor proteins that extrudes loops (reviewed in **Introduction**). Much less is known about how cohesins function in *C. elegans*, especially in the absence of CTCF-like proteins. *C. elegans* has singular homologs for Scc3, Smc1, and Smc3, but four for Scc1/Rad21 - COH-1, COH-2/SCC-1, COH-3, and REC-8. There is evidence that each of these Scc1 homologs regulate different processes. RNA interference (RNAi) of SCC-1 interferes with mitotic segregation and that of REC-8 causes defect in diakinesis during meiosis (Mito *et al* 2003). However, RNAi of COH-1 results in embryonic or larval stage arrests without mitotic or meiotic dysfunction (Mito *et al* 2003), suggesting a non-canonical and separate function for COH-1.

There are three condensin complexes in *C. elegans* - condensin I, condensin II, and a condensin-I-like dosage compensation complex that only differs from canonical condensin I by one subunit. Condensin I and II bind similarly in interphase and are enriched at active promoters (Kranz *et al* 2013), but have different chromosomal localisation in mitosis and meiosis (Csankovszki *et al* 2009,

Collete *et al* 2011), condensin II having a more dominant role during mitosis. Only subunits of the condensin II complex (MIX-1 [shared with condensin DC], SMC-4, KLE-2, HCP-6, and CAPG-2) were enriched at both interaction ends.

Condensins are not as well-studied as their Structural Maintenance of Chromosomes (SMC) family counterpart, but there are *in vitro* evidence that the *S. cerevisiae* condensin complex translocates along DNA (Terakawa *et al* 2017) and extrudes loops (Ganji *et al* 2018). In mice, clusters of condensin and transcription factor IIIC (TFIIIC) complexes are enriched at boundary-boundary interactions (**Fig 3.4**: blue arrows) (including non-adjacent boundaries) (Yuen & Gerton *et al* 2018), hinting at a role for these complexes in mediating domains. In a condensin II subunit knockdown experiment, most of the affected genes were located at these boundaries (Yuen *et al* 2017). The expression and interaction between histone gene clusters at TAD boundaries also require condensin (Yuen *et al* 2017). Cohesins and condensins appear to have complementary capacity for organising domains in mammals.

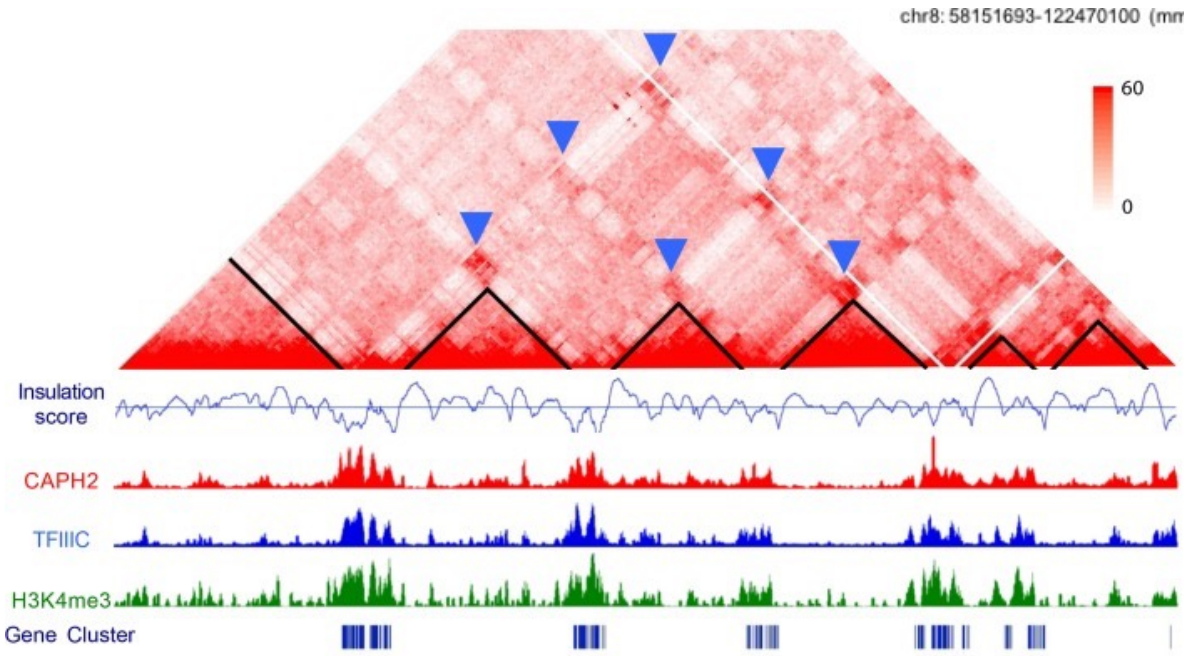
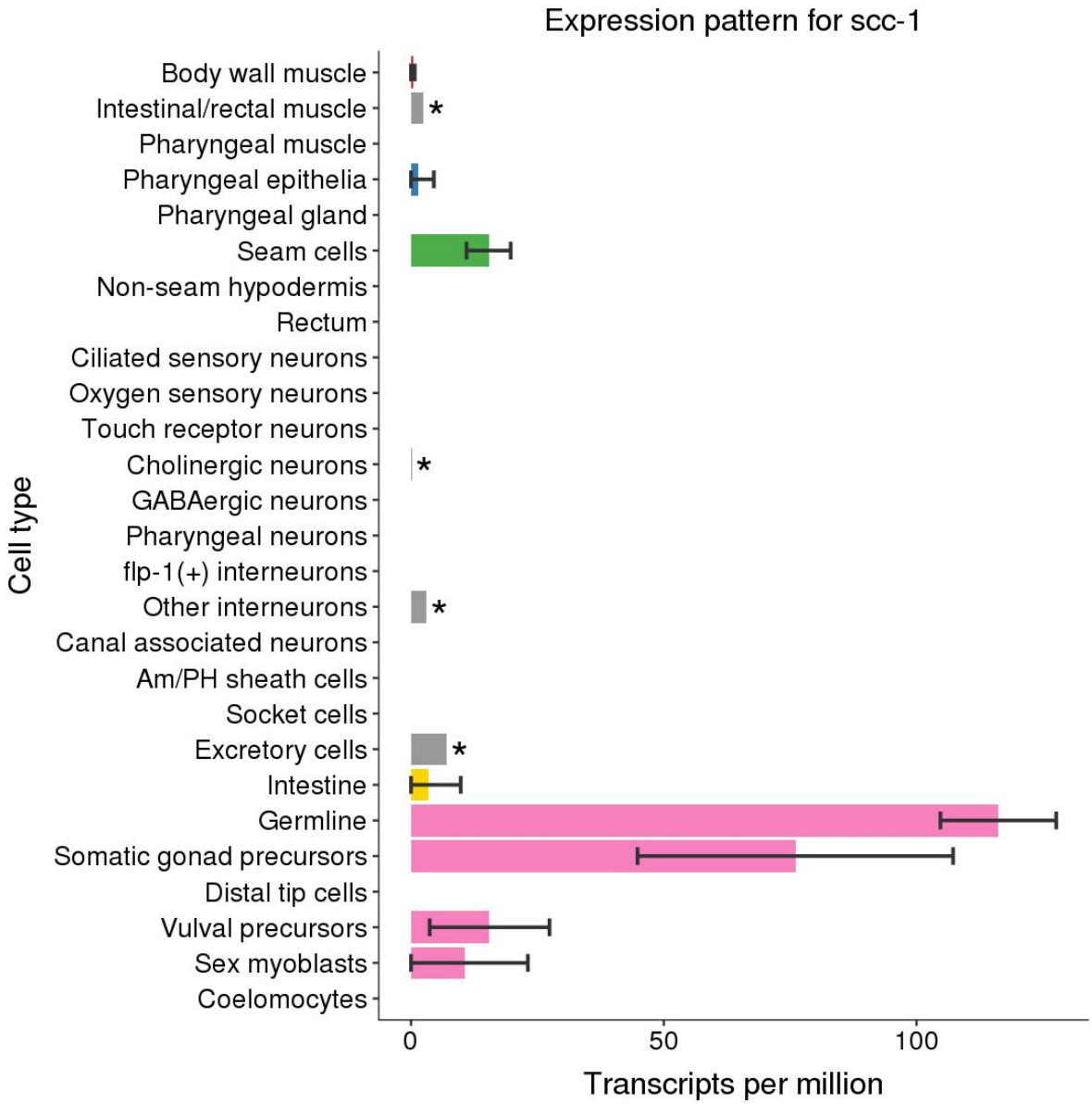


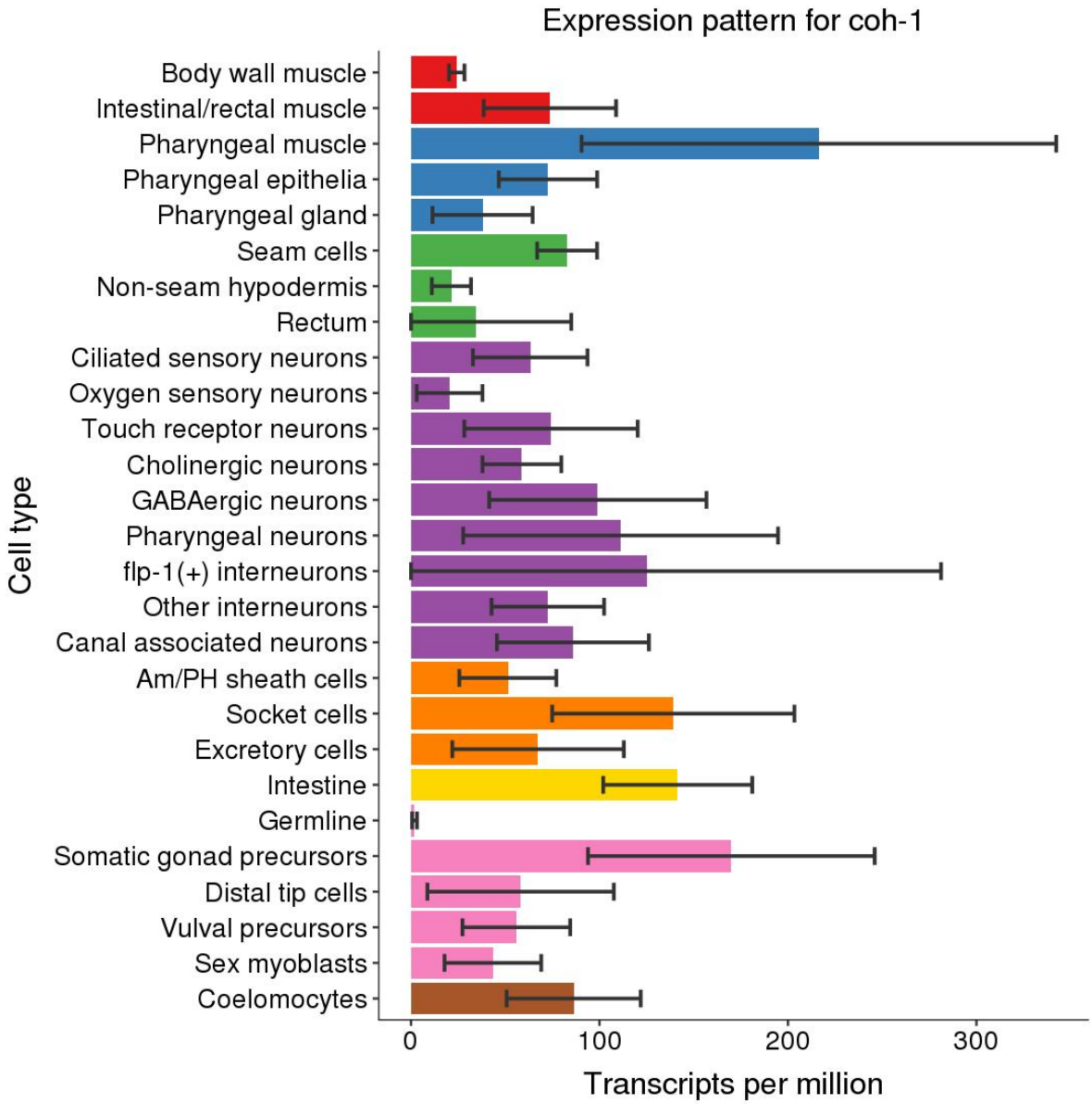
Figure 3.4: Snapshot of contact map in mouse (chr8: 58,151,693-122,470,100) (Yuen *et al* 2018). Top to bottom: contact map, insulation score, CAPH2 (condensin II) ChIP, TFIIIC ChIP, H3K4me3 ChIP, genes.

To look at the relationship between cohesin and condensin II, I compared their binding profiles. SCC-1 (cohesin), COH-1 (cohesin), and KLE-2 (condensin II) are expressed in different tissues. SCC-1 is highly enriched in the germline, KLE-2 is in the germline and soma, and COH-1 is soma-specific (**Fig 3.5**). SCC-1 and KLE-2 share similar binding characteristics (**Fig 3.6**; r = 0.68), which suggests similarities in function, while COH-1 has a relatively unique binding pattern genome-wide (**Fig 3.6**; KLE-2/COH-1, $r = 0.16$; SCC-1/COH-1, $r = 0.04$). I next clustered KLE-2, COH-1, and SCC-1 by k-means centred at regulatory elements. While COH-1 is distributed across all regulatory elements and dispersed across the 2kb window tested, SCC-1 and KLE-2 are strongly enriched in only about 15% of the elements (Clusters C2 & C4) (**Fig 3.7**). Interestingly, these two clusters had

high overlaps with HOT regions: 381/509 (95.481%, Fisher's exact test: $p < 0.0001$) for Cluster C2 and 2035/2754 (73.893%; $p < 0.0001$) for Cluster C4; 77.260% for Clusters C2 and C4 combined (and 14.804% for Clusters C1, C3, and C5; Fisher's exact test: n.s.). Given the structural similarities as SMC complexes and common binding loci, it could be plausible that SCC-1 and KLE-2 share similar functions, albeit in different tissues.

Interesting, there are clusters of SCC-1 and KLE-2 that bind adjacent to regulatory elements (**Fig 3.7** - C3 & C5). 45.3% of the regulatory elements centred in C3 and C5 had other regulatory elements with SCC-1 and KLE-2 binding within 1kb away in the corresponding direction (i.e. upstream in C3 and downstream in C5) at a median distance of 535bp. It is plausible that in the remaining cases, SCC-1 and KLE-2 are recruited adjacent to regulatory elements and the directionality of their binding parallels observation of CTCF directionality in mammals.


Expression pattern for scc-1
Body wall muscle
Intestinal/rectal muscle
Pharyngeal muscle
Pharyngeal epithelia
Pharyngeal gland
Seam cells
Non-seam hypodermis
Rectum
Ciliated sensory neurons
Oxygen sensory neurons
Touch receptor neurons
Cholinergic neurons
GABAergic neurons
Pharyngeal neurons
flp-1(+) interneurons
Other interneurons
Canal associated neurons
Am/PH sheath cells
Socket cells
Excretory cells
Intestine
Germline
Somatic gonad precursors
Distal tip cells
Vulval precursors
Sex myoblasts
Coelomocytes
Cell type
0
50
100
Transcripts per million


Expression pattern for coh-1
Body wall muscle
Intestinal/rectal muscle
Pharyngeal muscle
Pharyngeal epithelia
Pharyngeal gland
Seam cells
Non-seam hypodermis
Rectum
Ciliated sensory neurons
Oxygen sensory neurons
Touch receptor neurons
Cholinergic neurons
GABAergic neurons
Pharyngeal neurons
flp-1(+) interneurons
Other interneurons
Canal associated neurons
Am/PH sheath cells
Socket cells
Excretory cells
Intestine
Germline
Somatic gonad precursors
Distal tip cells
Vulval precursors
Sex myoblasts
Coelomocytes
Cell type
0
100
200
300
Transcripts per million

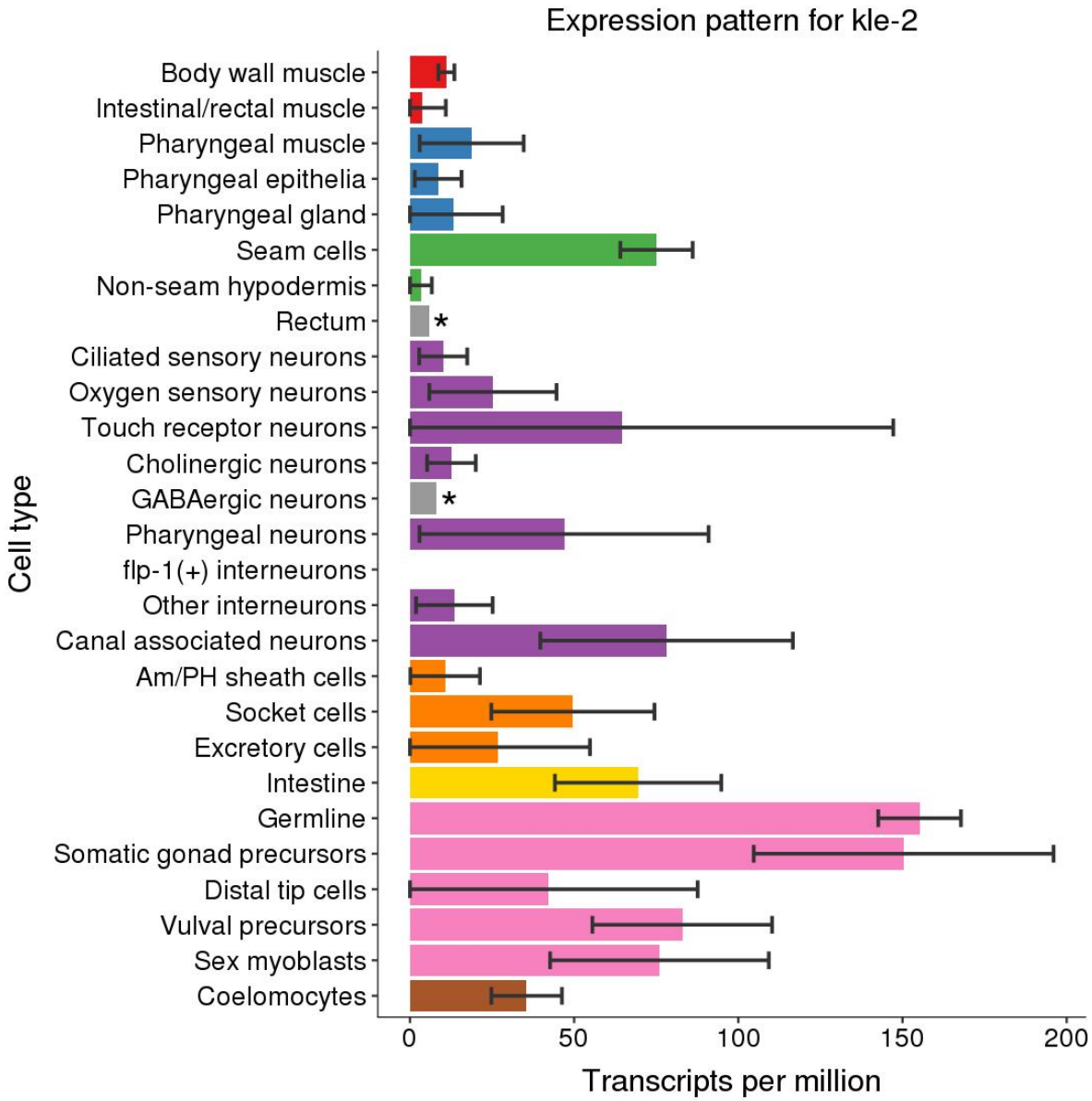


Figure 3.5: Expression profiles of *scc-1* (top), *coh-1* (middle), and *kle-2* (bottom) in different tissues.

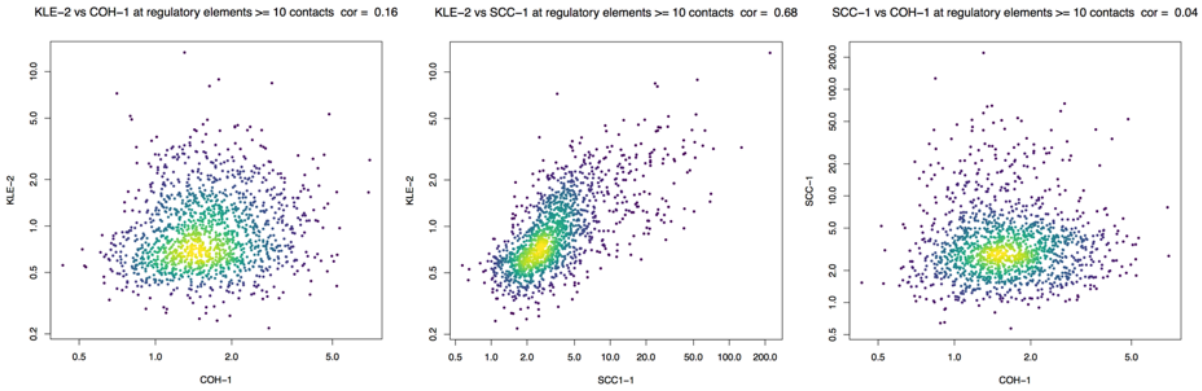


Figure 3.6: Correlation of KLE-2, COH-1, SCC-1 ChIP-seq at regulatory elements with at least 10 interactions. KLE-2 vs COH-1 (left, $r = 0.16$), KLE vs SCC-1 (middle, $r = 0.68$), SCC-1 vs COH-1 (right, $r = 0.04$).

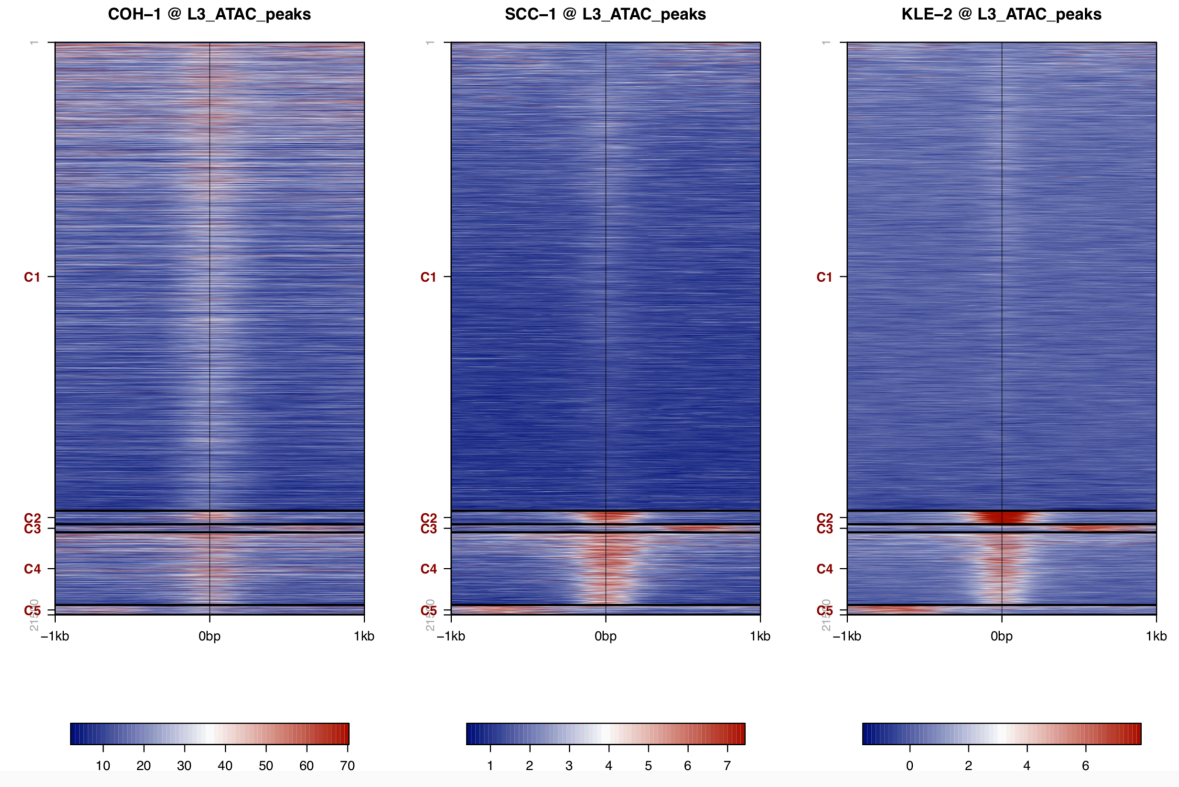


Figure 3.7: Heatmap of COH-1 (left), SCC-1 (middle), KLE-2 (right) in 2kb windows centred over L3 ATAC peaks.

COH-1 forms focal peaks that are aligned with SCC-1 and KLE-2 (**Fig 3.8**: an example is shown in the black box), but its binding is also dispersed around these peaks (**Fig 3.7** & **Fig 3.8**). These broad spreads of COH-1 binding frequently occur within active chromatin state domains ($p < 0.001$) and contrariwise, are depleted in regulated domains, as shown in an aggregate coverage plot over pseudo-scaled active or regulated domains (**Fig 3.10**). Intriguingly, stretches of COH-1 binding are also associated with clusters of significant interactions (**Fig 3.9**). Taken together, these associations raise the interesting possibility that COH-1 could be involved in the creation of clusters of interactions by carrying loci and translocating across DNA within active domains as a cause or consequence of gene activity. These results are also consistent with observations of condensin clusters and their relation to dense clusters of interactions and gene activity (inferred through H3K4me3 binding) as discussed earlier (**Fig 3.4**) (Yuen & Gerton *et al* 2018), implicating COH-1 in domain formation.

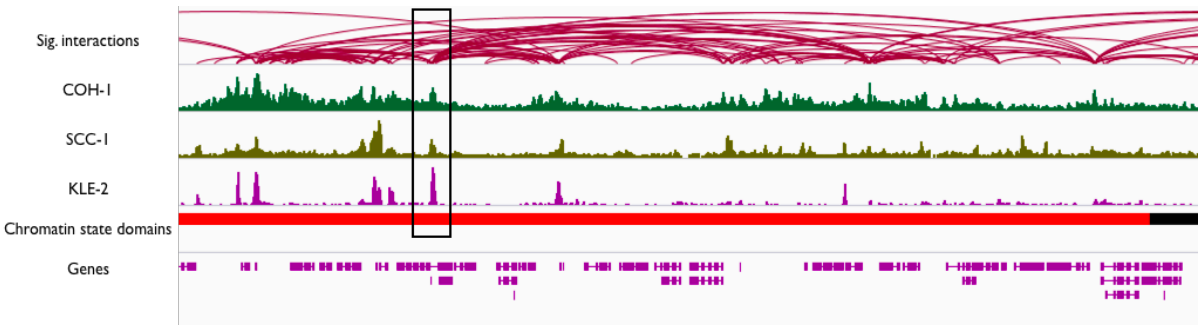


Figure 3.8: Snapshot of (top to bottom) significant interactions, COH-1 ChIP, SCC-1 ChIP, KLE-2 ChIP, chromatin state domains - active (red), regulated (black), genes. Black box indicates shared focal peaks. ChrI: 7450,441-7,612,266.

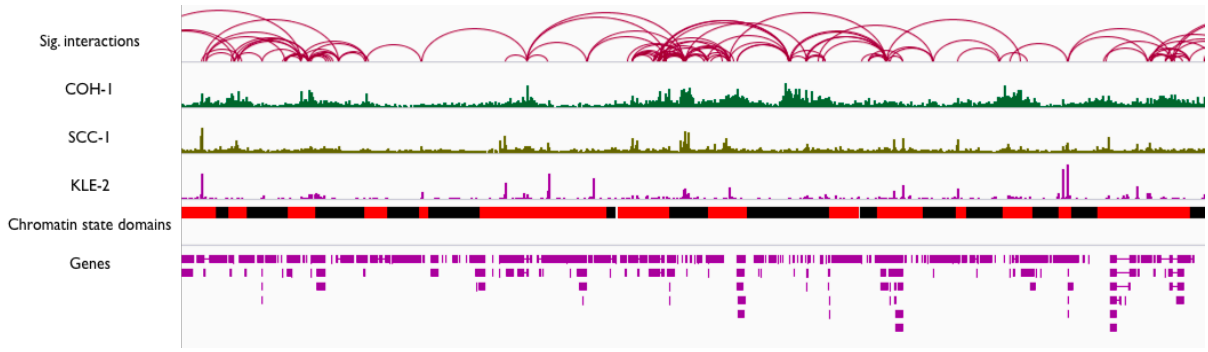


Figure 3.9: Snapshot of (top to bottom) significant interactions, COH-1 ChIP, SCC-1 ChIP, KLE-2 ChIP, chromatin state domains - active (red), regulated (black), genes. COH-1 clusters overlap significant interaction clusters

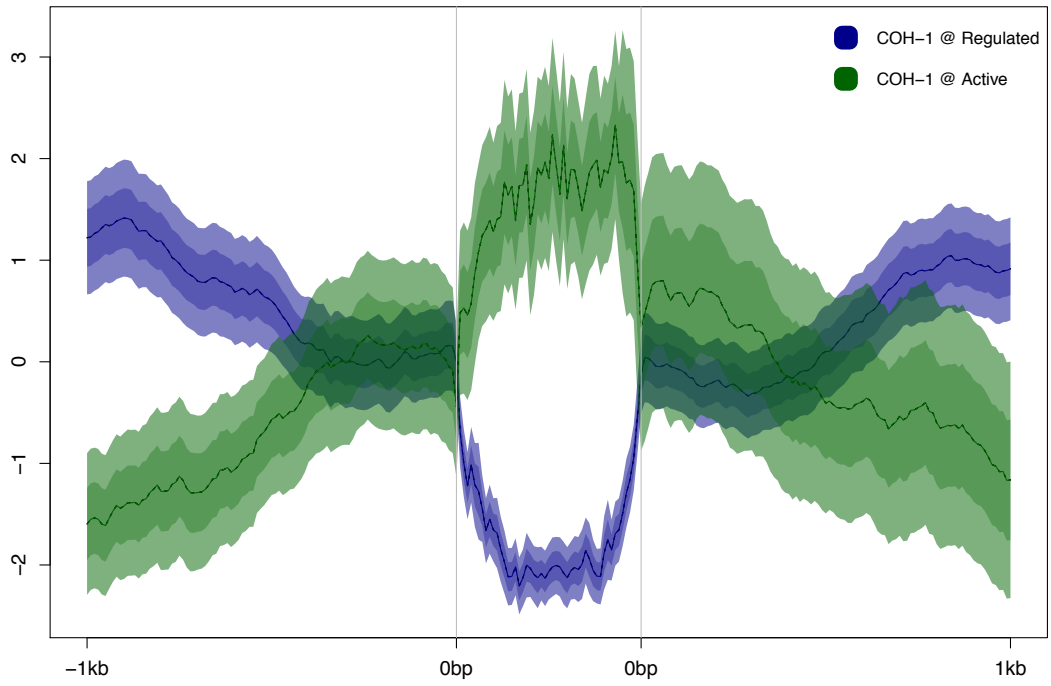


Figure 3.10: Aggregate plot of z-scored COH-1 ChIP signal over pseudo-scaled active (green) or regulated (blue) chromatin state domains.

Chromatin regulators

We also identified components from several chromatin remodelling and regulating complexes, such as the SIN3 histone deactylase, Nucleosome Remodelling and Deacetylase (NuRD), Nucleosome Remodelling Factor (NURF), Dimerisation Partner, RB-like, E2F and Multi-vulval class B (DREAM), and AF10/ DOT-1 (**Table 3.2**). Notably, a majority of these have known repressor activity. While they have not been extensively studied in the context of 3D genome architecture, it is unsurprising that they could be implicated. NuRD or NURF interact with cohesin to regulate access to nucleosomal DNA during gene regulation and chromosome segregation. In fact, in humans, SCC1/RAD21 was reported to physically contact SNF2h, the ATPase subunit of NuRD (Hakimi *et al* 2002). DNA binding to CTCF-bound Alu repeat sequences requires SNF2h (Hakimi *et al* 2002; Fasulo *et al* 2012). In mice, NURF binds CTCF-target sites *in vivo* and its subunit Bptf has physical interactions with CTCF and cohesin through SA2/SCC3 (Qiu *et al* 2015). At insulator sites in *Drosophila*, the DREAM complex homolog (dREAM) potentially co-operate with NURF to open up heterochromatin (Le Gall *et al* 2015) and the enhancer-blocking ability of insulator sites (synonymous with TAD boundaries) require both dREAM and NURF (Bohla *et al* 2014). These studies indicate that a multitude of chromatin remodellers and regulators co-operate - through shared subunits or other transcription co-

regulators - to regulate chromatin interactions. It is plausible that some of the factors we have identified cooperate to control chromatin interactions.

To determine which factors have potential to work together, I asked if any of them had significantly overlapping binding sites. I calculated Jaccard index scores for pairs of chromatin regulator ChIP-seq peak-sets as a measure of similarity and implemented an unsupervised hierarchical clustering to identify groups of factors that have more shared binding sites. As a control, I did the same for the highly correlated condensin subunits (Albritton *et al* 2018) and obtained scores from 0.104 to 0.355 with a median of 0.229 (**Fig 3.11**).

In general, binding sites for chromatin regulators were correlated; Jaccard indices ranged from 0.0201 to 0.515 with a median of 0.152 (**Fig 3.12**). As expected, related subunits cluster together: SIN3 (SIN-3, EPC-1, HCF-1), DREAM and NURD (LIN-35, LIN-53), and AF10/DOT1 (ZFP-1, DOT-1.1). Interestingly, AF10/ DOT1 appears to share binding sites with the NuRD subunit LIN-40 (0.196 with ZFP-1 and 0.240 with DOT-1.1) and SIN3 correlates with the NuRF subunit NURF-1. Put together, our factor APA have implicated certain chromatin in loop formation; evidence from other studies and our binding analyses suggest they could collaborate and further studies are required to elucidate this relationship and potential mechanisms of action.

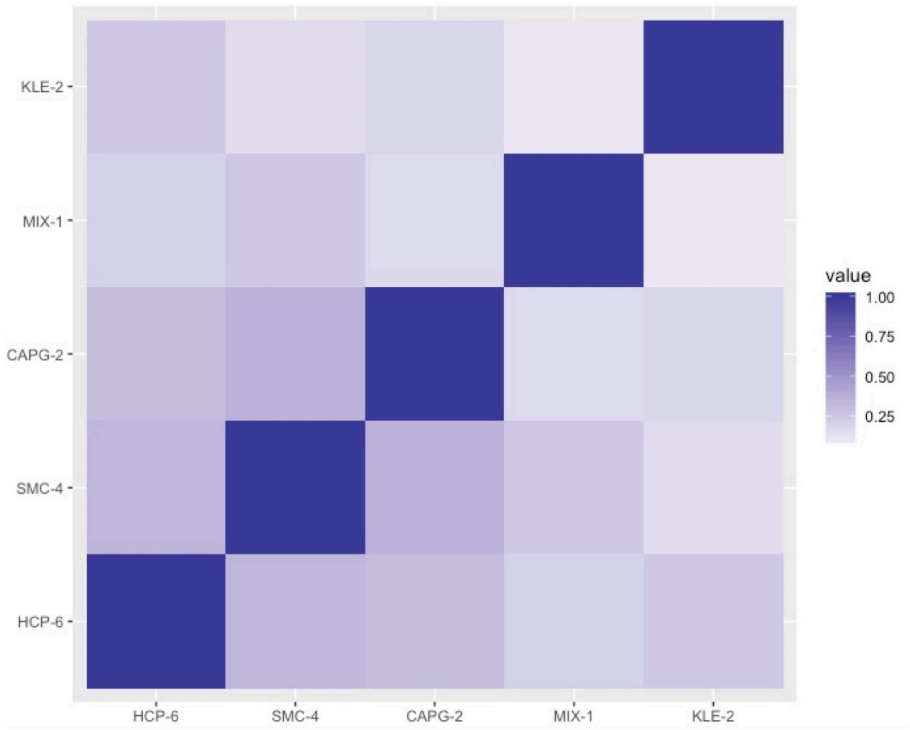


Figure 3.11: Jaccard index heat map of condensin II subunits.

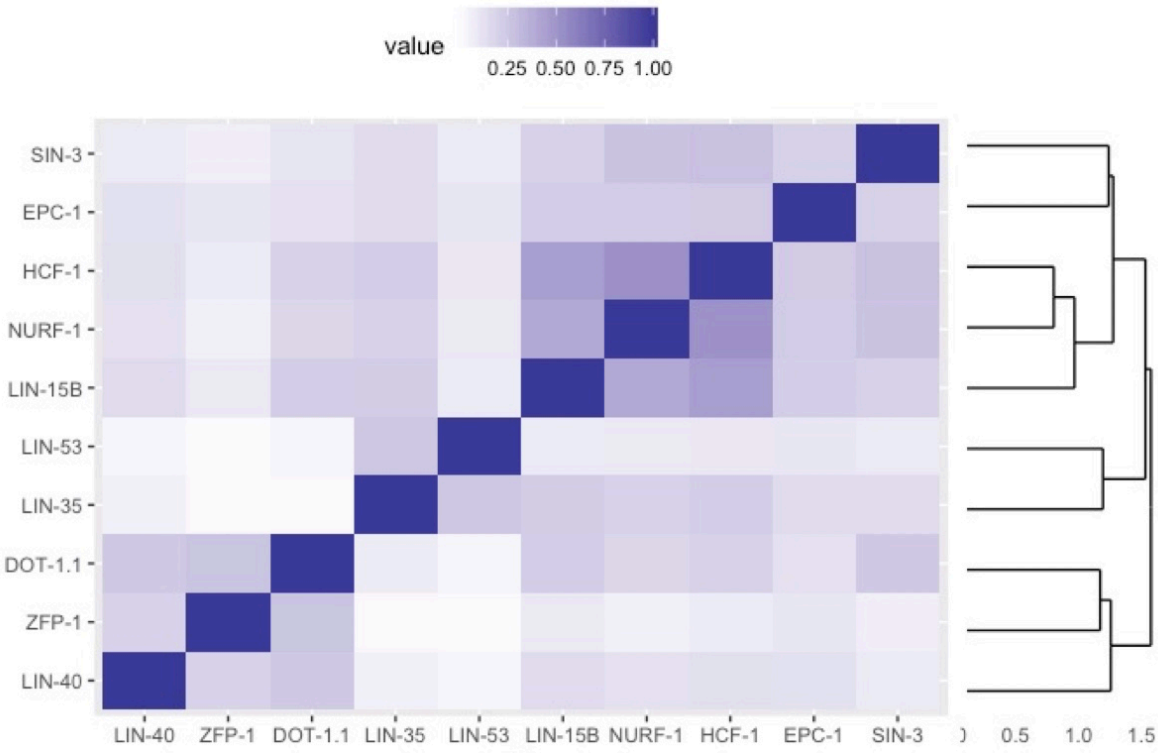


Figure 3.12: Jaccard index heat map of chromatin regulators identified by factor APA.

Transcription factors

The factor APA also identified various transcription factors that were expressed or enriched in major tissues in L3 stage larvae, namely the muscles (ELT-2, FOS-1, HLH-1, NHR-28, RNT-1), hypodermis (BLMP-1, ELT-2, ELT-3, FOS-1, NHR-23, NHR-28, RNT-1), and intestine (BLMP-1, ELT-2, ELT-3, FOS-1, NHR-28, NHR-80, PHA-4, PQM-1). Unfortunately, given that ARC-C was done in whole animals, it is difficult to interpret the biological significance of this assemblage of factors - i.e, why some factors were identified but not others. Such an understanding might be forthcoming if ARC-C was performed in homogeneous samples.

ARC-C in *blmp-1* mutants

I wanted to test the effects of depleting a factor from our list of 'hits' on chromatin looping. I selected the transcription factor BLMP-1 from the list because unlike many chromatin remodellers, *blmp-1* mutants grow relatively synchronously up to L3 stage, have observable non-lethal mutant phenotypes, and is one of the factors that are crucial to L3 stage chromatin accessibility (Daugherty *et al* 2017). I selected the allele *tm548* which contains a 810 bp deletion removing parts of exon 3 and intron 3, resulting in a truncated and ostensibly non-functional protein (Huang *et al* 2014).

BLMP-1, the ortholog of the mammalian zinc finger transcription repressor BLIMP-1/PRDM-1, cooperates with LIN-40/MTA-1 to regulate stress responses and downstream targets such as *nhr-23* and *sams-1* (Hyun *et al* 2016), a S-adenosyl methionine synthase. However, there is also evidence that BLMP-1 can function as a transcription activator under hypoxia by activating stress-response pathways (Padmanabha *et al* 2015). BLMP-1, a heterochronic gene, also interacts with DRE-1 to control larval molting, dauer formation, epidermal and gonadal development (Horn *et al* 2014; Huang *et al* 2014). *blmp-1* mutants experience incomplete alae formation, a weakly retarded developmental phenotype (Horn *et al* 2014). In my hands, when grown at 20C, *blmp-1* (*tm548*) grew slower (36 h to reach mid-L3

stage), were dumpy and had a weakly penetrant embryonic lethality (~5 to 10% dead eggs each generation). I collected *blmp-1* mutants at the mid-L3 stage and performed ARC-C, generating roughly 7.2 million *cis*, informative reads (**Appendix - Table A1.1**).

We wanted to investigate if the loss of BLMP-1 led to any changes in chromatin looping at regions under the influence of BLMP-1 binding regions. First, we defined "BLMP-1 targets" as BLMP-1 peaks that overlap annotated regulatory elements of genes misregulated in L3 stage *blmp-1* as compared to N2 based on poly(A)-RNA-seq. 822 genes were upregulated and 736 genes were downregulated ($0.667 <$ fold-change < 1.5, adjusted $p < 0.05$) in *blmp-1* mutants. The gain or loss of chromatin loops could also affect genes locally by modulating physical proximity. We thus defined "BLMP-1 near targets" as BLMP-1 peaks within 2 kb of misregulated genes, regardless of their annotation status. Factor APA was then conducted using *blmp-1* ARC-C data with the same windows used in the N2 factor APA earlier. Additional APAs were done for BLMP-1 targets and BLMP-1 near targets, separated by all, up-, and downregulated genes.

blmp-1 ARC-C had much lower signal over DHS than N2 (1.94 to 2.75 in *blmp-1* vs. 3.7 to 5.5 in N2) (**Table 1.6** & **Appendix - Table A1.1**) and direct comparisons between both strains could produce artefactual results. To control for

this, we compared BLMP-1 APA with the other factors. Representative APAs for a few factors (ELT-2, KLE-2, and SCC-1) are shown in **Fig 3.13**. As expected, APA scores in *blmp-1* were consistently lower than in N2 (within 14.06% to 25.39% for the non-BLMP-1 factors tested), but substantially different for BLMP-1 (49.57%) (**Fig 3.13**).

This consistent decrease in APA score is reflected in the positive trend we observed when we plot the log2 of factor APA scores in *blmp-1* against the same in N2 (**Fig 3.14**), which we modelled linearly and through a locally weighted scatterplot smoothing (LOWESS) (**Fig 3.15** and **Fig 3.16**). However, neither methods are sufficiently robust at the moment as the confidence intervals are based on several relatively arbitrary factors that do not evenly and unbiasedly sample all potential interactions. We are currently working on a more statistically rigorous way for testing significance.

BLMP-1 targets and BLMP-1 near targets appeared to be outliers from the general trend but they were mainly driven by the downregulated genes in both instances (**Fig 3.15** and **Fig 3.16**). Upregulated BLMP-1 targets and near targets did not deviate much from the trend-line (**Fig 3.15** and **Fig 3.16**). Possible reasons could be that the repressor function for BLMP-1 is looping-independent, that the loss of BLMP-1-mediated interactions were compensated by activating factors

creating ‘new’ or maintaining ‘old’ loops, or that upregulated genes were an indirect effect of BLMP-1 knockout. Consistent with studies that showed that the loss of loops is more correlated to a reduction in transcriptional activity than the gain of loops is with transcriptional activation (Phanstiel *et al* 2017), the loss of interactions (presumably BLMP-1-mediated) at BLMP-1-target genes was associated with a significant downregulation in gene expression. Further work is required to explore the relationship between factor-binding, loop formation, and gene expression, but preliminary results in *blmp-1* mutants suggest that it is involved in formation of loops that have transcriptional consequences.

The factor APA in ARC-C identified factors that were enriched at both ends of interactions. Part of these - cohesin, condensin, NuRD - have been implicated in 3D genome organisation in other organisms. Others relate to transcription factors and chromatin remodellers that can be further tested. The transcription activation function of one of the transcription factor, BLMP-1, appears to be sensitive to a loss in chromatin interactions. These results pave the way for further studies in the other identified factors. That said, the limitations of factor APA are that it primarily tests for factors underlying loops, not domains or compartments (which I discuss in the next chapter), requires the use of factors that produce narrow ChIP peaks, and is dependent on the quality of ChIP-seq data used.

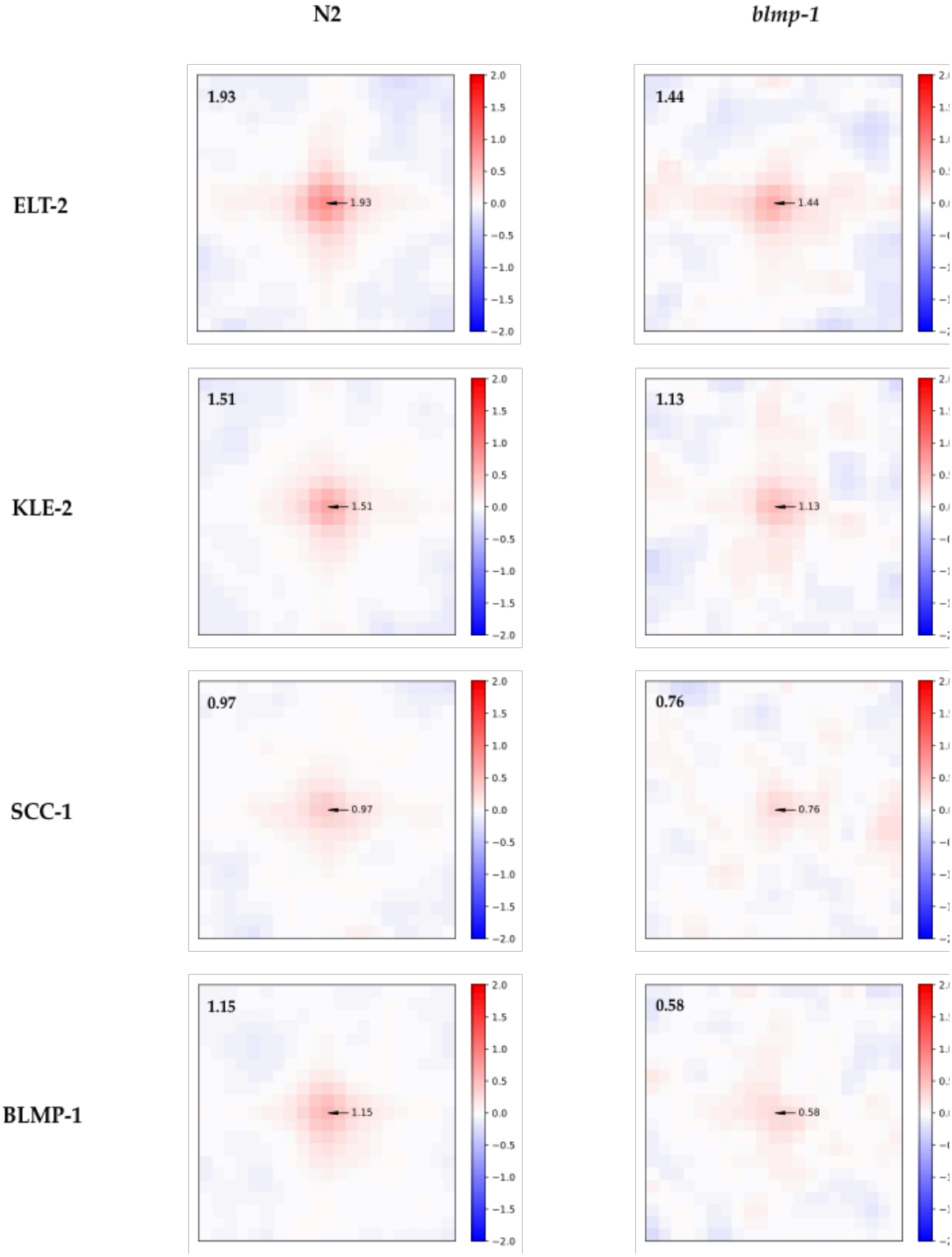


Figure 3.13: Factor APA of BLMP-1 and control factors ELT-2, KLE-2, SCC-1 in N2 and *blmp-1* mutants at 1kb resolution. Numbers indicate APA log2FC over background of corresponding factors.

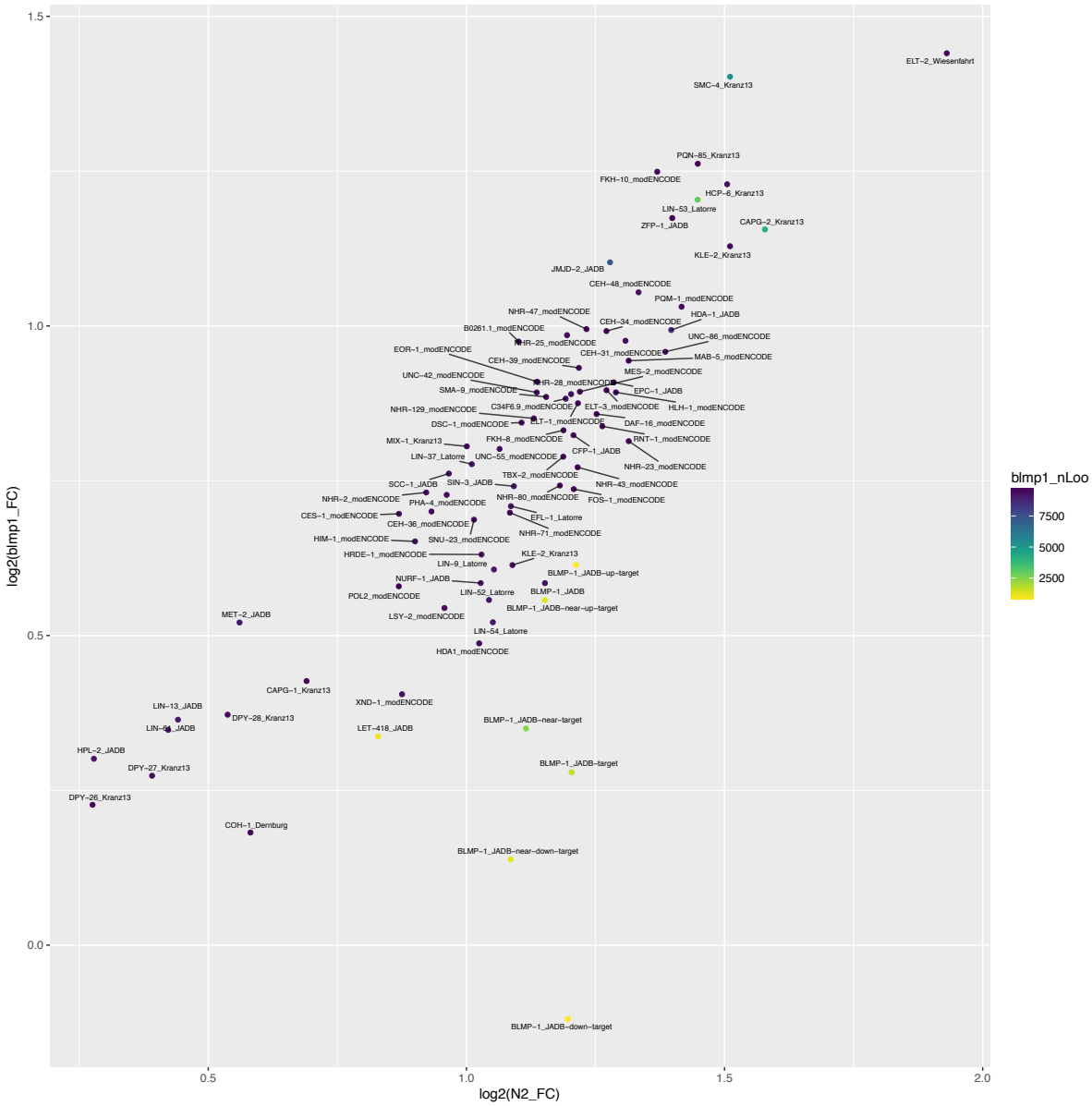


Figure 3.14: Log2-log2 plot of factor APA scores in wild-type (N2) over *blmp-1* mutants. The number of loops indicates the number of paired peaks (or APA windows) used.

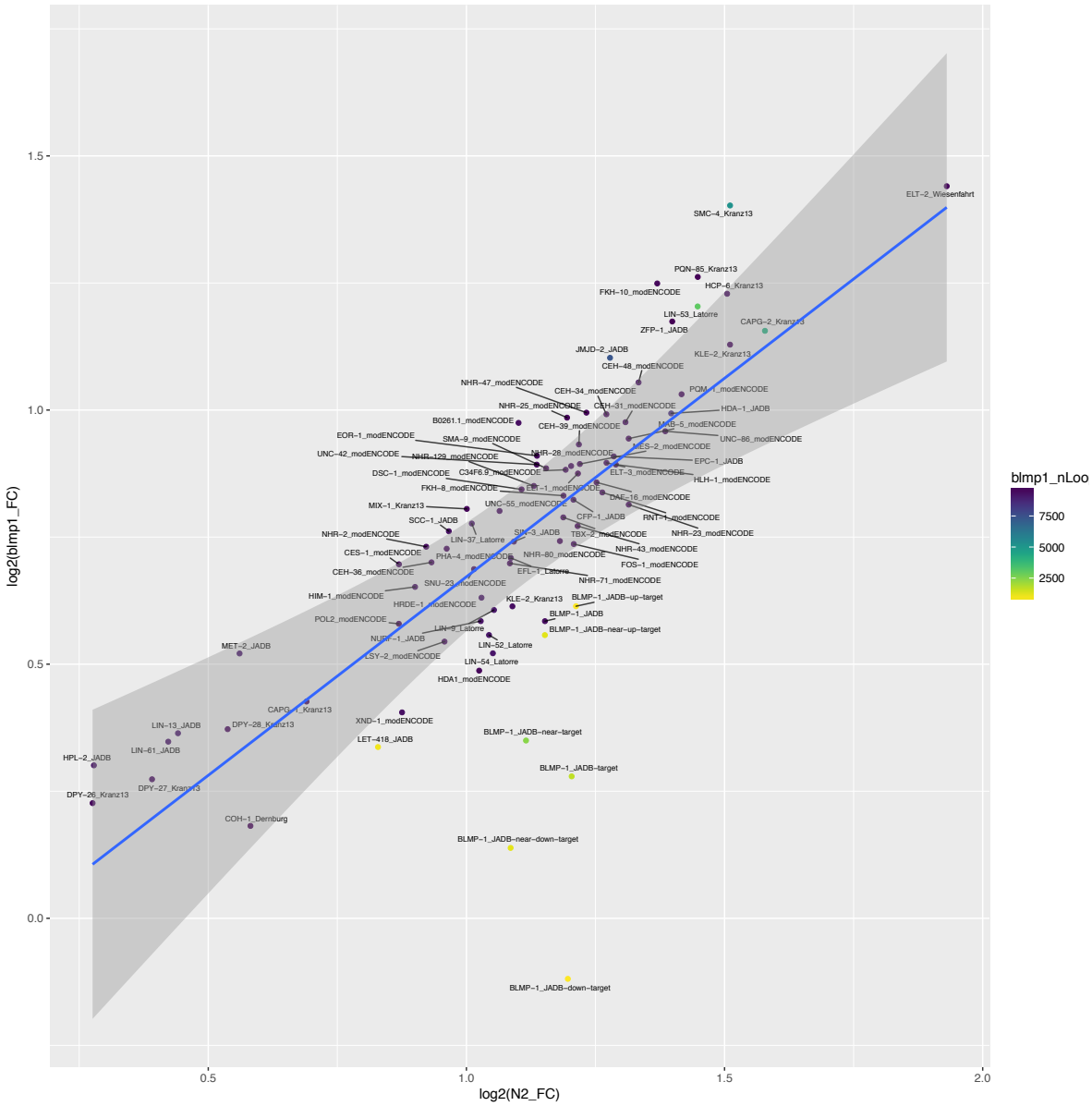


Figure 3.15: Log2-log2 plot of factor APA scores in wild-type (N2) over *blmp-1* mutants. The number of loops indicate the number of paired peaks (or APA windows) used. The positive trend is modelled linearly.

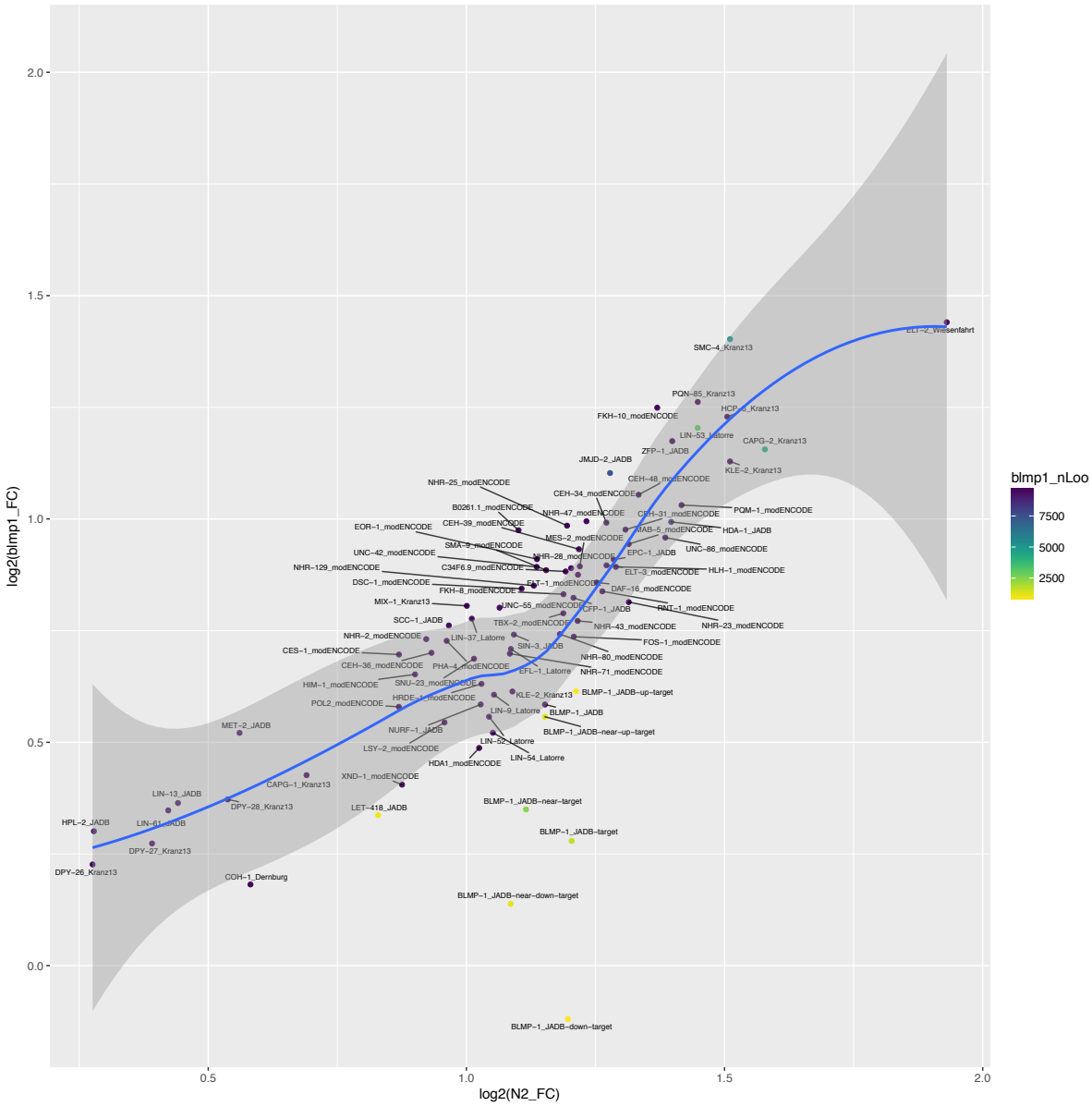


Figure 3.16: Log2-log2 plot of factor APA scores in wild-type (N2) over *blmp-1* mutants. The number of loops indicate the number of paired peaks (or APA windows) used. The positive trend is modelled with LOWESS.

CHAPTER IV: DOMAINS AND COMPARTMENTS

In Crane Hi-C, TADs as defined by local contact insulation were not detected on the autosomes (Crane *et al* 2015). Likewise, neither directionality index (Dixon *et al* 2012) nor contact insulation (Crane *et al* 2015) produced sensible TAD calls in ARC-C (data not shown). In *Drosophila*, TADs correlate well with epigenetic domains, namely active domains of H3K4me3 and H3K36me3, PcG repressed domains enriched with H3K27me3 and inactive chromatin (Sexton *et al* 2012, Ulianov *et al* 2016). Repressed TADs form nanocompartments when visualised with 3D-structured illumination microscopy (Szabo *et al* 2018). In *C. elegans*, we report for the first time the existence of TAD-like domains and compartments on the autosomes.

Previous work in the lab segmented the autosomal chromatin into 20 chromatin states, but these states can be broadly summarised into H3K36me3-enriched active chromatin and H3K27me3-enriched regulated chromatin (Evans *et al* 2016). H3K27me3 forms broad stretches along the *C. elegans* genome (median length = 18,670 bp, n = 1,257) and can be summarised into domains that we term "regulated" (**Fig 4.1**). These domains are associated with inactive and lowly expressed genes, as well as developmentally and conditionally regulated genes

(Evans *et al* 2016). H3K9me3 is also known to co-occur with H3K27me3, particularly at the chromosome arms (Ho *et al* 2014). Regulated domains alternate with other domains that we term "active" (**Fig 4.1**). Active domains (median length = 13,608 bp, n = 1,274) are associated with active genes in the highest quintile of expression and mark genic elements such as promoters, enhancers, and transcription elongation, and are also enriched for H3K36me3 (Evans *et al* 2016).

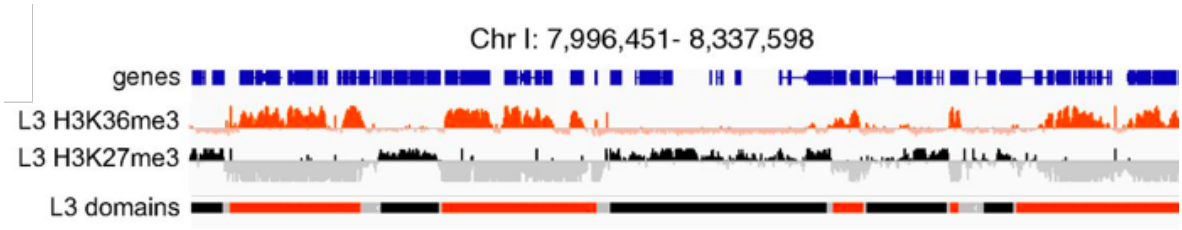


Figure 4.1: Representative view of H3K36me3 and H3K27me3 forming broad expanses, which were summarised into active (orange) or regulated (black) domains (Evans *et al* 2016).

Domains and compartments in wild-type worms

Given the absence of insulators in *C. elegans* and compact genome, I hypothesised that epigenetic domains could form weakly insulating contact domains or compartments and tested for the presence of these by aggregate plots - domain APA or compartment APA. To survey for TAD-like domains, we looked along the diagonal of the contact matrix (illustrated in **Fig 4.2**: Domain APA) while centring on the chromatin state domains. To see if these domains form compartments, we looked at the off-diagonal space and focused on all possible inter-domain permutations (illustrated in **Fig 4.2**: Compartment APA). These

windows were then pseudoscaled and overlaid; the contact frequency of these aggregated domains were compared with their neighbourhood (**Fig 4.3**: adjacent blocks illustrated by yellow dotted lines) to measure the extent of insulation for domain APA and compartmentalisation for compartment APA as compared to background. An intensely shaded block in these plots would indicate insulation or compartmentalisation.

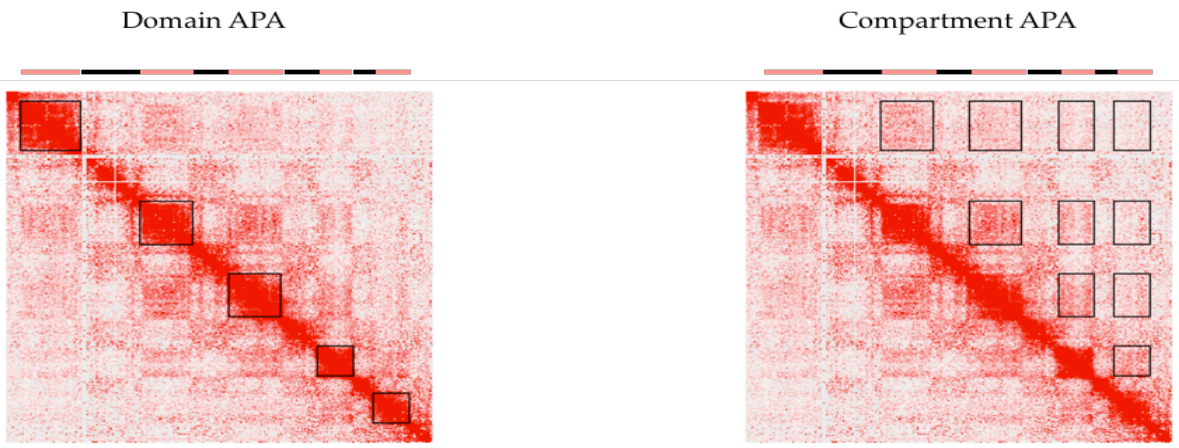


Figure 4.2: Illustration of Domain APA and Compartment APA. Pink bars represent regulated domains; black bars represent active domains. Boxes indicate the windows used in a domain (left) or compartment (right) APA of regulated domains.

We investigated the ability of both regulated and active domains to form insulated contact domains and compartments by performing APA at 5 kb resolution, doing so for the arms and centre regions separately and together, and at different distance intervals for compartments. Active and regulated domains do

establish contact domains and compartments. Active domains form stronger contact domains and compartments than regulated domains at all instances, reflecting the increased chromatin activity associated with active domains (**Fig 4.3**). This phenomenon is not due to differences in the number of genes in either domains; both domain-types have the same median number of genes per domain (3) (Evans *et al* 2016).

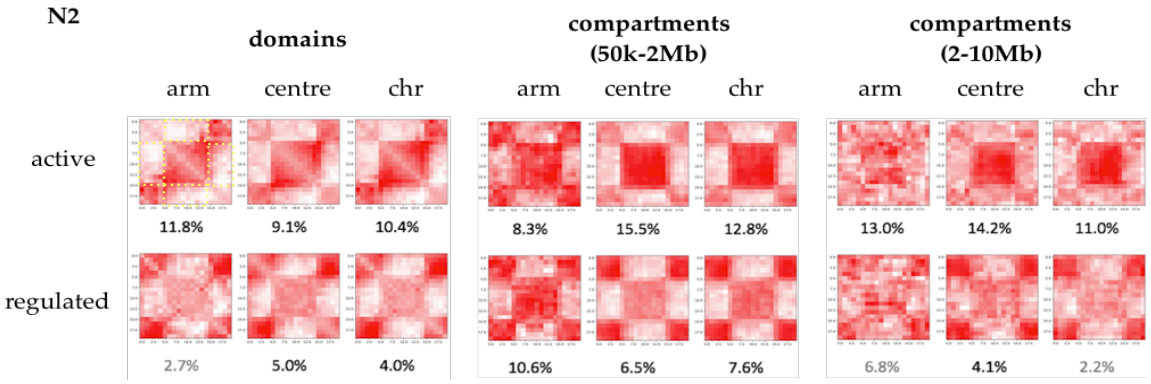


Figure 4.3: Domain and compartment APA (50kb-2Mb, 2-10Mb) for active and regulated chromatin state domains in wild-type ARC-C. Yellow dotted boxes indicate the neighbourhoods that the centre blocks are compared against. The numbers measure the percentage enrichment of the centre blocks over neighbourhoods.

Active compartments are stronger at the centre than at the arms at both short (50k-2Mb) and long range (2-10Mb), with this effect being greater for short-range compartments (6.6% stronger at short range, $p < 0.01$; 1.1% at long range, $p > 0.5$) (**Fig 4.3**). This is likely due to the centre region being more transcriptionally active than the arms, as mentioned previously. The reverse is true for regulated

compartments - they have a higher contact frequency at the arms, suggesting that domains and compartments are potentially formed or maintained by separate mechanisms. Finally, interactions between regulated domains are relatively local and are lost at larger distances as compared to interactions between active domains, which remain strong at short and long distances (**Fig 4.3**). This implies that there might be robust, active (or non-equilibrium) mechanisms maintaining these long range interactions in active compartments (these will be explored further in subsequent chapters). Active and regulated linear chromatin state domains, as defined by H3K36me3 and H3K27me3 marks respectively, form insulated contact domains and compartments. These domains and compartments can also be found in Hi-C in aggregate (data not shown), validating their existence in *C. elegans*.

We next compared compartments in *C. elegans* with those in *Drosophila* to better understand and appreciate the compartment APA scores we obtained earlier. The fly was selected as it shares a comparably small, compact genome. Compartments in *Drosophila* were defined as clusters of TADs that were post factum defined as active or inactive based on epigenetic features (Sexton *et al* 2012). We performed compartment APA with the active or inactive TADs defined in Sexton *et al* 2012 at 50kb to 2Mb and 2Mb to 10Mb (**Table 4.4**). Within the distance intervals tested, the *Drosophila* embryo genome possesses stronger levels

of compartmentalisation (17.5-22.5% in *Drosophila* vs 2.2-12.8% in the worm) (**Table 4.4**). This is unlikely due to the use of mixed-tissue samples for *C. elegans*, given that chromatin state domains that were separately defined in early embryos and L3 stage larvae were fairly consistent (Evans *et al* 2016). In the paradigm of loop extrusion, the prevalence of insulator proteins which are enriched at TAD boundaries in *Drosophila* (Sexton *et al* 2012, Hou *et al* 2012, Stadler *et al* 2017) and the lack thereof in *C. elegans* cannot explain stronger compartmentalisation in the fly as insulators or loop extrusion barriers are thought to be antagonistic to compartmentalisation (Nuebler *et al* 2018): the removal of CTCF in mouse suppresses TADs but leaves compartments relatively pristine (Nora *et al* 2017). Compartmentalisation is also distance-dependent and weakens with increasing distance in all instances (**Table 4.4**), peaking at 500 kb for both active and inactive clusters in *Drosophila* and 200 kb for both active and regulated domains in *C. elegans* when we further subdivide the distance ranges (data not shown). In contrast to reports that did not find compartments in *C. elegans* (Crane *et al* 2015), we show here that compartments do exist, although they are marginally weaker than in *Drosophila*.

Sample	Compartment type	Distance interval	Source	Compartment APA (%)
***Drosophila* Oregon-R w^{1118} 16-18 h embryos (Hi-C)**	Active	50 kb - 2 Mb	Sexton *et al* 2012	19.6
		2 Mb - 10 Mb		17.5
	Inactive	50 kb - 2 Mb		22.5
		2 Mb - 10 Mb		20.5
***C. elegans* N2 L3 stage larvae (ARC-C)**	Active	50 kb - 2 Mb	-	12.8
		2 Mb - 10 Mb		11.0
	Regulated	50 kb - 2 Mb		7.6
		2 Mb - 10 Mb		2.2

Table 4.4: Summary of compartment APA scores in *Drosophila* and *C. elegans* in active and inactive/regulated domains at 50kb-2Mb and 2-10Mb.

Role of H3K9 methylation in domains and compartments

As discussed in the **Introduction**, heterochromatin can form compartments through the self-association of heterochromatic proteins. In contrast to *Drosophila* and human, H3K9me3 and H3K27me3 have significant overlaps (Ho *et al* 2014), particularly at chromosome arms. This co-occurrence suggests a level of cooperativity in function, given also that both marks are necessary for transgene silencing (Towbin *et al* 2012). There is evidence that H3K9methylation (H3K9me) affects chromatin architecture. H3K9me3 is associated with LEM-2 domains, an inner nuclear membrane associated with A-type lamins (Ikegami *et al* 2010), suggesting H3K9me3 might be involved in nuclear periphery localisation.

Subsequently, Towbin *et al* (2012) showed that heterochromatic attachment of transgene arrays to the nuclear envelope requires H3K9me.

The different states of H3K9me relate differently between human, fly, and worm. H3K9me3 is less correlated with H3K9me2 in the worm (r = 0.40) as compared to fly (r = 0.89), while H3K9me2 is well correlated with H3K9me1 in the worm (r = 0.44) but slightly anti-correlated in the fly (r = -0.32) (**Fig 4.5**; Ho *et al* 2014), which suggests H3K9me2 could function differently to H3K9me3 in the worm as compared to the human and fly. This raises questions about the role of different H3K9 methylation states in compartmentalisation.

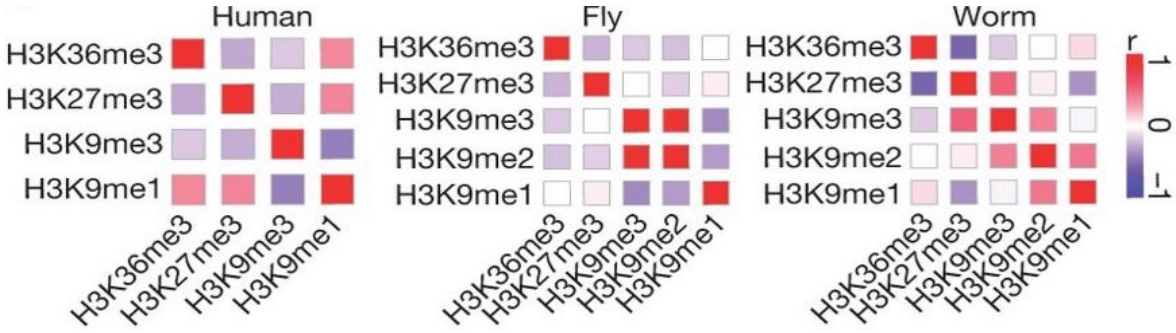


Figure 4.5: Correlation of histone marks across human, fly, and worm (Ho *et al* 2014).

H3K9me2 and H3K9me3 are enriched on the chromosome arms (**Fig 4.6**: e.g. chr I). To define H3K9me2 and H3K9me3 domains for downstream analyses, we called broad peaks using MACS2 and built a model for such domains using Hidden Markov Modelling (HMM) at 1 kb resolution with short segments under 5 kb filled in with neighbouring states; a representative example is shown in **Fig 4.7** for H3K9me2 and H3K9me3. Both H3K9me2 and H3K9me3 domains had significantly high overlaps (at least 70% reciprocal overlap) with regulated domains: 441/536 (82.28%, $p < 0.001$) for H3K9me2 and 403/459 (87.80%, $p < 10^{-7}$) for H3K9me3. Despite similar median lengths (12 kb for H3K9me2 and 13 kb for H3K9me3), H3K9me2 domains also intersected active domains (at least 70% reciprocal overlap) more frequently than H3K9me3 domains (**Fig 4.7**): 356/536 (66.42%) for H3K9me2 and 58/459 (12.64%) for H3K9me3. We then performed domain and compartment APAs with these H3K9 methylation domains with wild-type ARC-C data. Results were shown alongside the active and regulated chromatin state domains APAs as in **Fig 4.3** but with a different colour scale (**Fig 4.8**).

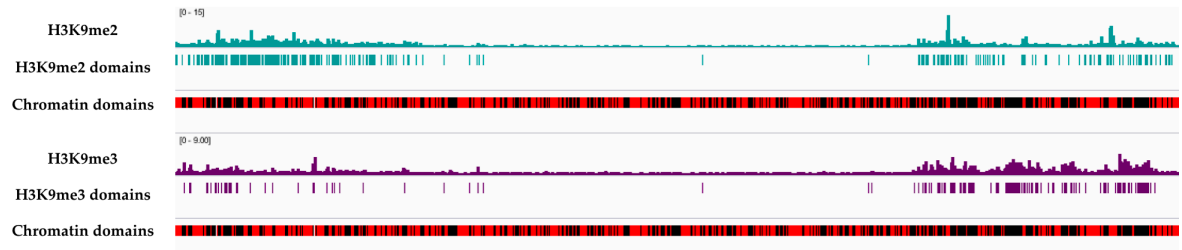


Fig 4.6: Top to bottom: Distribution of H3K9me2/3 over chr I. HMM-called H3K9me2/3 domains. Active (red) and regulated (black) chromatin state domains. Chr I.

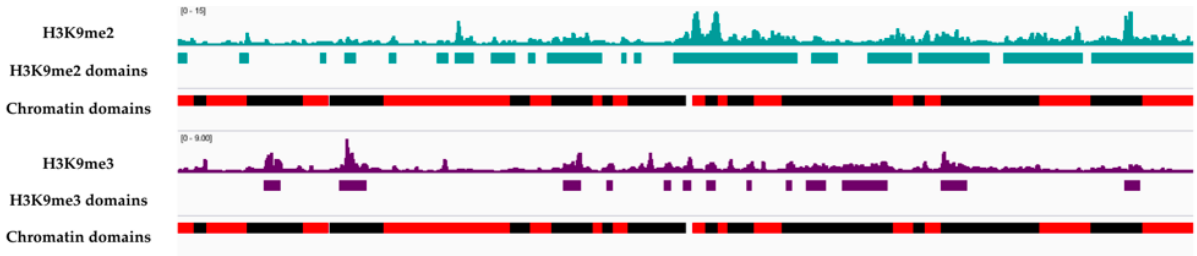


Fig 4.7: Zoomed in view of **Fig 4.6**; chr I: 100,000 - 1,100,000.

In terms of relative strengths for contact domains, active-arm domains were the strongest; regulated-arm domains were comparable to H3K9me3 domains, which were both stronger than H3K9me2 domains (**Fig 4.8**). Regulated-arm domains form compartments as strong as in active-arm and H3K9me3 domains (**Fig 4.8**). It is likely that the tethering of the arm, more specifically at LEM-2 domains that correspond to H3K9me3 binding (Ikegami *et al* 2010), contributes to this phenomenon. Similar to its ranking for contact domains, H3K9me2 domains produce the weakest compartments, indicating that H3K9me2 is likely not involved in regulated-domain compartmentalisation. In all, the results reveal a putative role for H3K9me3 but not H3K9me2 in the formation or maintenance of regulated domains and compartments.

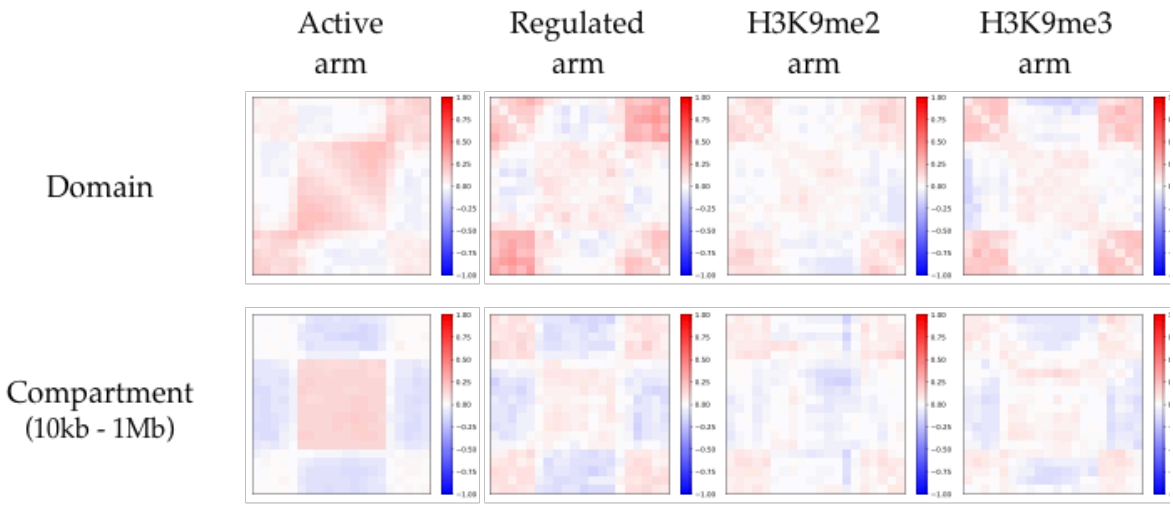


Fig 4.8: Domain and compartment APA for active, regulated, H3K9me2, H3K9me3 domains in wild-type ARC-C.

Domains and Compartments in *met-2 set-25* mutants

Having established that H3K9 methylation might have a role in regulated compartments, we set out to validate that by doing ARC-C in H3K9 methylation-deficient worms. This question is particularly relevant given experiments on inverted nuclei and polymer modelling by Falk *et al* (2018) that suggest heterochromatic interactions are pivotal to the phase separation of active and inactive chromatin or compartmentalisation, while euchromatic interactions are dispensable for this process. *met-2* (*n4256*) *set-25* (*n5021*) double mutants (strain GW637) lack mono-, di-, and trimethylated H3K9 (Towbin *et al* 2012, Garrigues *et al* 2015), with cytoplasmic MET-2 catalysing mono- and dimethylation (Andersen & Horvitz 2007, Bessler *et al* 2010) and nuclear SET-25 converting H3K9me1/2 to H3K9me3 (Towbin *et al* 2012). In *met-2 set-25* mutants, nuclear lamin binding to the chromosome arms was reduced and transgene GFP-reporter arrays lost

perinuclear localisation (Towbin *et al* 2012). However, *met-2* alone did not show array detachment while *set-25* mutants that lack H3K9me3 still had H3K9me1/2 enriched over the nuclear periphery and around 70% of the arrays still tethered (Towbin *et al* 2012). At 20C, *met-2 set-25* mutants grew normally with no obvious mutant phenotypes, albeit at a slightly slower rate (34 h to mid-L3 stage instead of 30 h). At 25C, they grew asynchronously, with low brood size and transgenerational sterility. I prepared ARC-C libraries from L3 stage *met-2 set-25* mutants grown at 20C and sequenced them to 10.7 million *cis*, informative reads (**Appendix - Table A1.1**).

Large overall changes in accessibility can affect how ARC-C data is interpreted. To test whether the loss of H3K9me affected accessibility, I performed ATAC-seq in *met-2 set-25* and compared signals with wild-type ATAC-seq at accessible sites defined in Jänes *et al* (2018) (described in **Chapter II**): coding/ pseudogene/unknown promoters, putative enhancers, ncRNA, and unannotated elements (**Fig 4.9**). Across the classes of regulatory elements, putative enhancers and unknown promoters had similar accessibility as wild-type worms (**Fig 4.9**). Coding promoters, pseudogene promoters, and ncRNA were significantly but not substantially (< 1.13-fold, $p < 0.001$) more accessible (**Fig 4.9**) and accordingly, 126 genes were upregulated and 9 genes were downregulated (adjusted $p < 0.05$) in *met-2 set-25* mutants. Accessibility was unchanged between wild-type and *met-2*

set-25 mutants for regulated domains but very marginally higher for active domains (1.037-fold) (**Fig 4.10**). In the top 10^{th} percentile of H3K9me2 binding in wild-type, gains in accessibility correlated with the levels of H3K9me2 binding (**Fig 4.11**); this was again a very small difference. In all, accessibility between wild-type and *met-2 set-25* worms were very similar, with the exception of a few regions highly bound by H3K9me2 and in active regions, and unlikely to confound domain and compartment analyses.

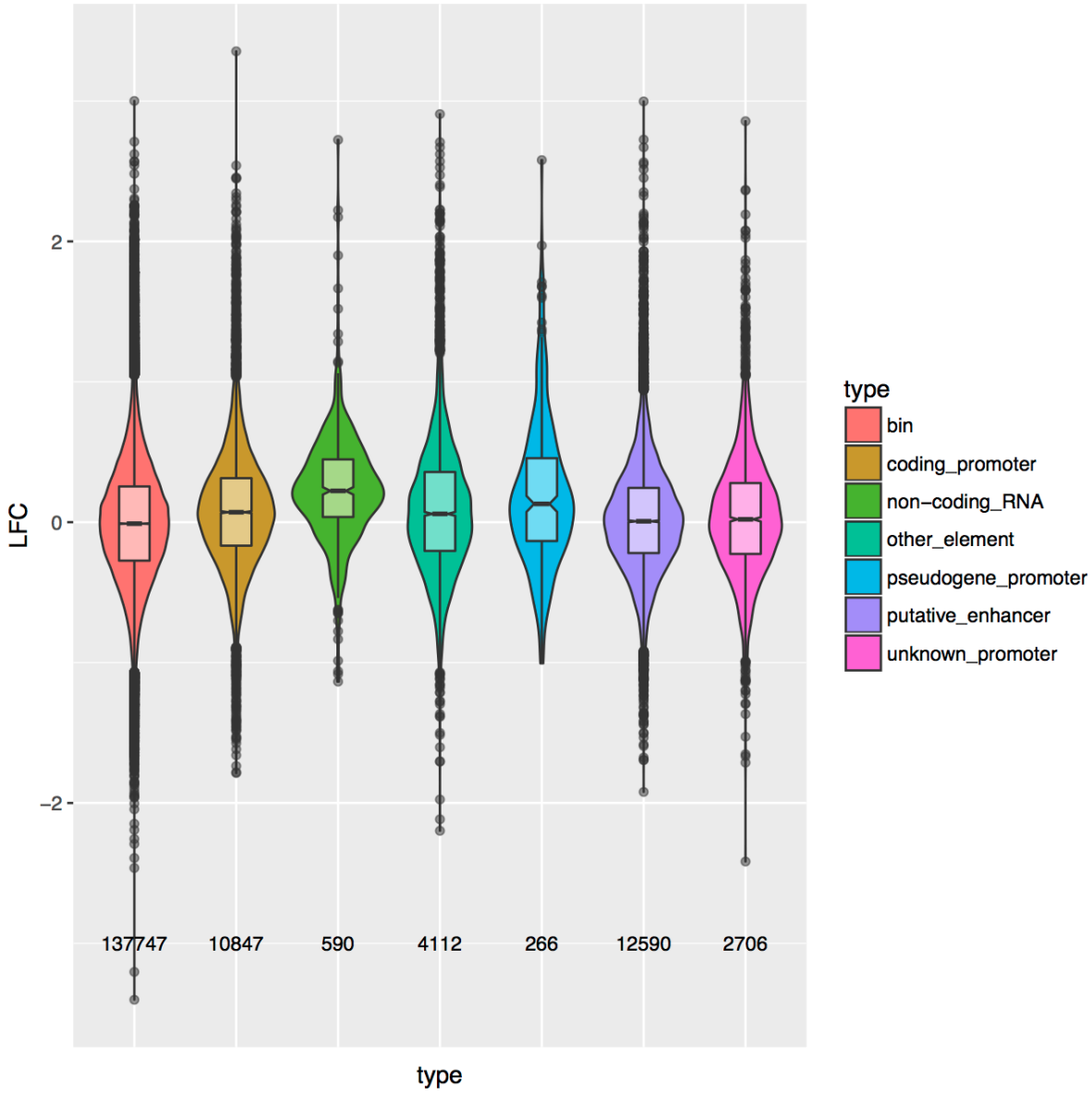


Fig 4.9: Log2FC in accessibility between *met-2 set-25* and wild-type worms, separated into all (red), coding promoters (orange), non-coding RNAs (green), unannotated elements (turquoise), pseudogene promoters (blue), putative enhancers (purple), and unknown promoters (magenta).

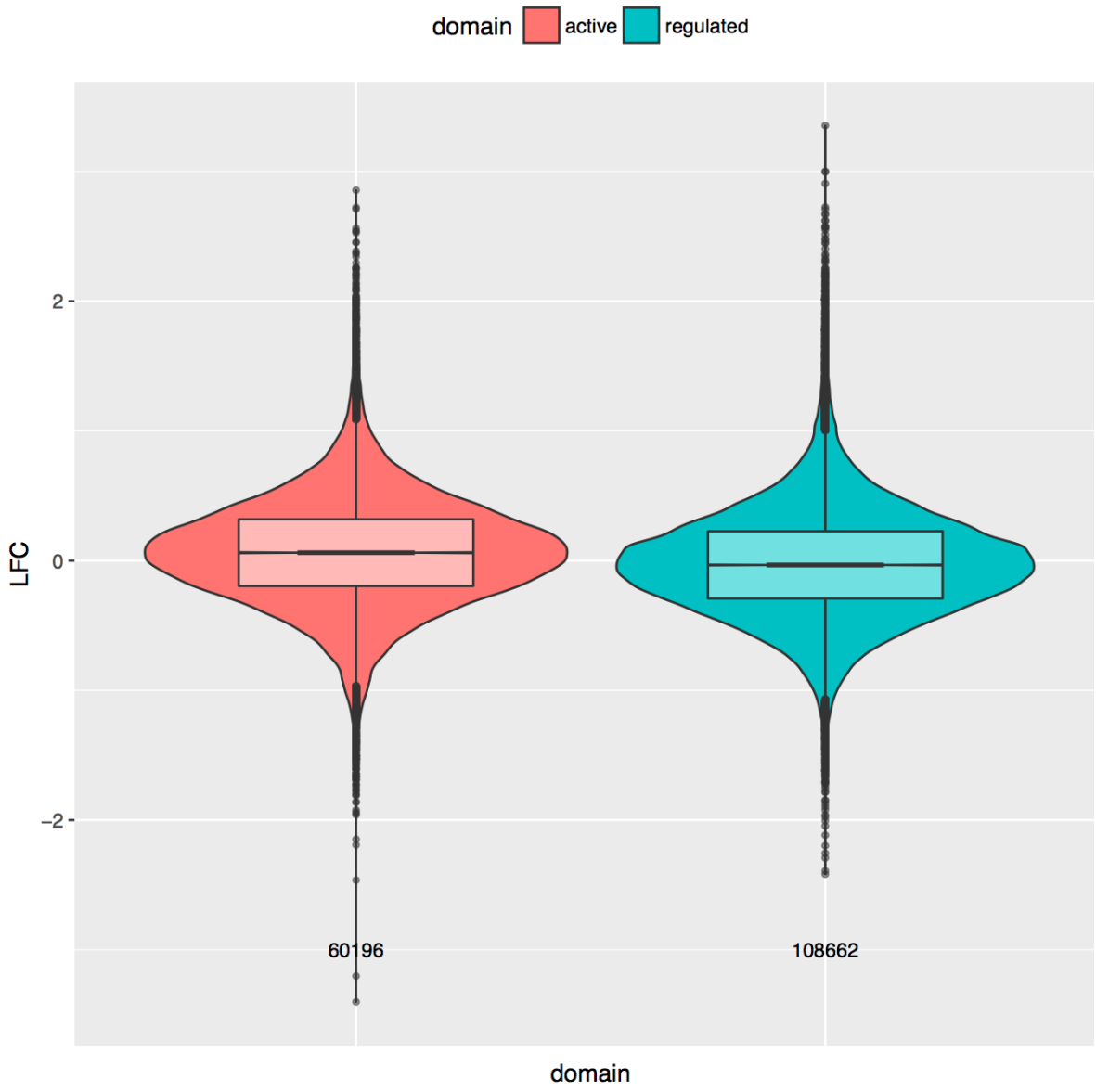


Fig 4.10: Log2FC in accessibility between *met-2 set-25* and wild-type worms within active (red) or regulated (green) chromatin state domains.

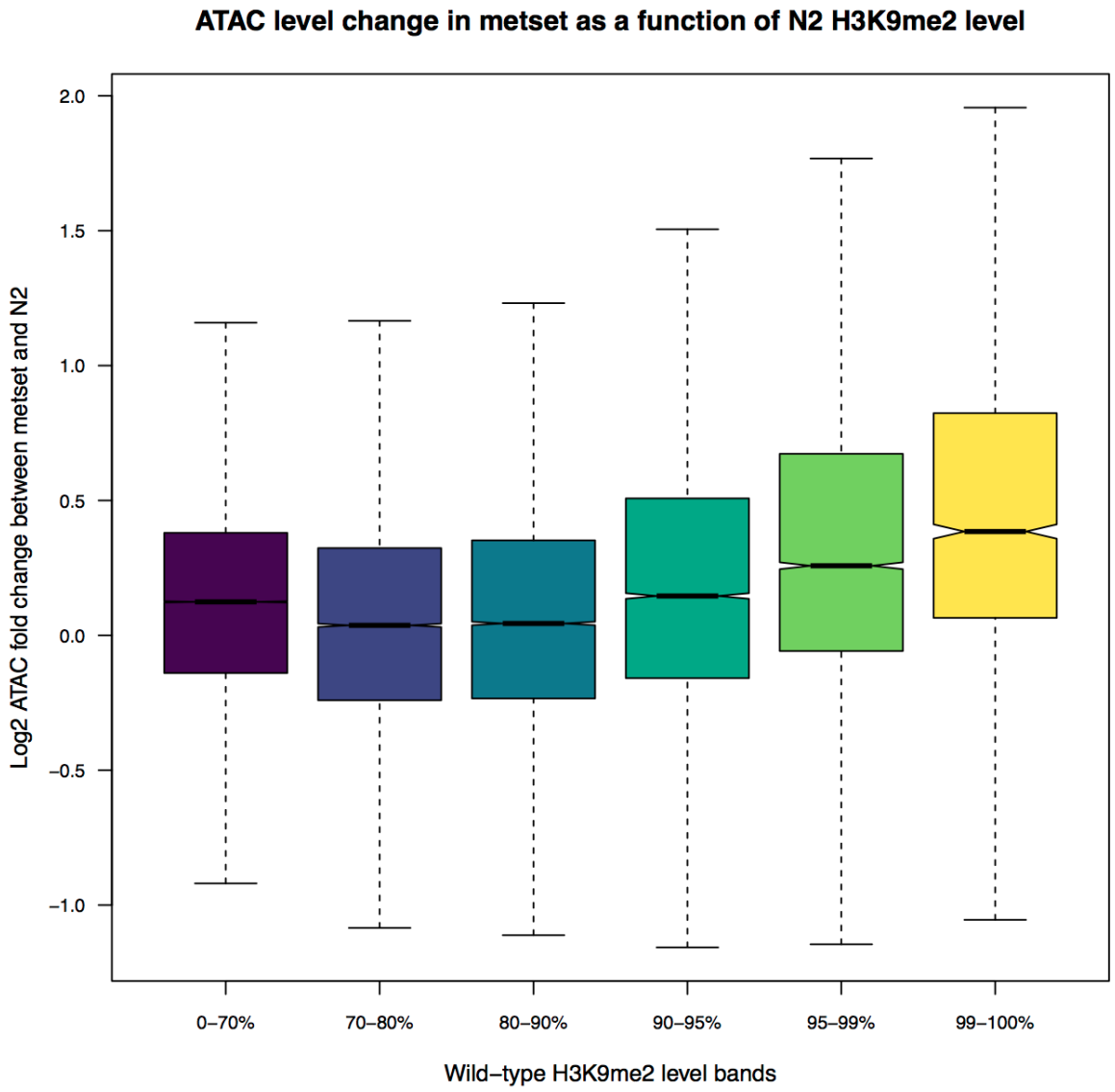


Fig 4.11: Log2FC in accessibility between *met-2 set-25* and wild-type worms at different levels of wild-type H3K9me2 binding.

Despite largely equivalent levels of overall accessibility, N2 and *met-2 set-25* mutants differed in their chromatin folding. The contact decay exponent describes the gradient of the slope with which contact frequency or contact probability diminishes over increasing distance from a particular frame of reference. In some studies, it is interpreted as a measure of local chromatin compaction. A steeper

slope implies long range interactions are less likely, which apparently translates to increased compaction (eg. Grob *et al* 2013, Grob *et al* 2014, Zhu *et al* 2017). The relationship between the exponent and physical compaction was corroborated by super-resolution microscopy (Boettiger *et al* 2016).

Indeed, this appears consistent when comparisons were made between heterochromatin and euchromatin (more negative exponent in heterochromatin; Zhu *et al* 2017), pericentromeric regions and euchromatin in *A. thaliana* (-1.243 vs -0.703; Grob *et al* 2014), and human and *Drosophila*/*Arabidopsis* genomes (-1.08 vs -0.85/-0.73; Grob *et al* 2013) where the former is thought to be much more compact. However, this supposition assumes homogeneity in compaction within the elements studied, which readily breaks down; aggregate measures of contact probability frequently and invariably overestimate the exponent as it ignores insulated, nested domains (Sanborn *et al* 2015). Exponents can vary within supposedly similar genomic and epigenetic environments (from -0.56 to -0.96 in Grob *et al* 2013). In fact, the relationship is inverted when comparing repressive Null, PcG, and HP1 domains with active domains in *Drosophila* (around -0.7 in Null/PcG/HP1 vs -0.85 in active; Sexton *et al* 2012). The appropriate interpretation at this moment when observing different exponents would be that there are separate interaction regimes.

We plotted the contact probability of valid interactions between 10 kb to 2.5 Mb in N2 and *met-2 set-25* mutants, split into each of the five autosomes and between the chromosome arms and central regions. In all instances (arm and centre), below 10kb, the *met-2 set-25* contact probability profile was steeper and above wild-type, suggesting an increase in short range interactions and a separate regime for chromatin packaging (**Fig 4.12**). On the arms, the exponents were alike from 10 kb up to approximately 500 kb but decreased and deviated subsequently for chr I, chr II, chr IV, and chr V (**Fig 4.12** & **Fig 4.13**). By contrast, in the central regions, the profiles matched up from 10 kb to around 500 to 750 kb and the exponents decreased considerably thereafter (**Fig 4.12** & **Fig 4.13**). Together, the data suggest *met-2 set-25* adopted different chromatin folding at very short (< 10kb) and very long (> 500kb) ranges.

As expected from their enrichment at the arms, *met-2 set-25* had a greater effect on the arms with a leftward translation from 100 kb to 2.5 Mb (**Fig 4.13**), implying the loss of long range interactions. The decrease in long range interactions and concomitant increment of short range interactions were corroborated by a plot of interaction frequency separated by distances (short: <20kb, median: 20-200kb, long: >200kb) and chromosome locations (**Fig 4.14**). The loss of H3K9me in *met-2 set-25* precipitated a change in chromatin organisation at

both ends of the distance distribution and implicated H3K9me in the mediation of long range interactions.

We next determined the effect of H3K9me loss on domains and compartments in aggregate. Regulated domains had fairly similar contact domains in *met-2 set-25* as compared to N2 (**Fig 4.15**). However, they had 1.058-fold weaker short range compartments (50kb - 2Mb) with the chromosome arms contributing a larger portion of the reduction (8% for arms and 5.4% for the central regions) (**Fig 4.15**). Long range (2 - 10Mb) regulated compartments in *met-2 set-25* were completely lost (**Fig 4.15**). Paradoxically, active domains formed weaker insulated domains (4.2%) but stronger short range compartments (3.9%) with no change in long range compartments (**Fig 4.15**).

One possible explanation could be that mechanisms for active and regulated interaction domain and compartment formation are antagonistic. Given the relatively short (~20kb) and alternating nature of active and regulated chromatin state domains, the weakening or loss of regulated domains in *met-2 set-25* could pave the way for neighbouring active domains to interact more strongly together, resulting in stronger short range compartments. Contrary to proposals that heterochromatic interactions are the sole driver of compartmentalisation (Falk *et*

al 2018), we find evidence that active and regulated compartments can form independently, with regulated compartments requiring H3K9me.

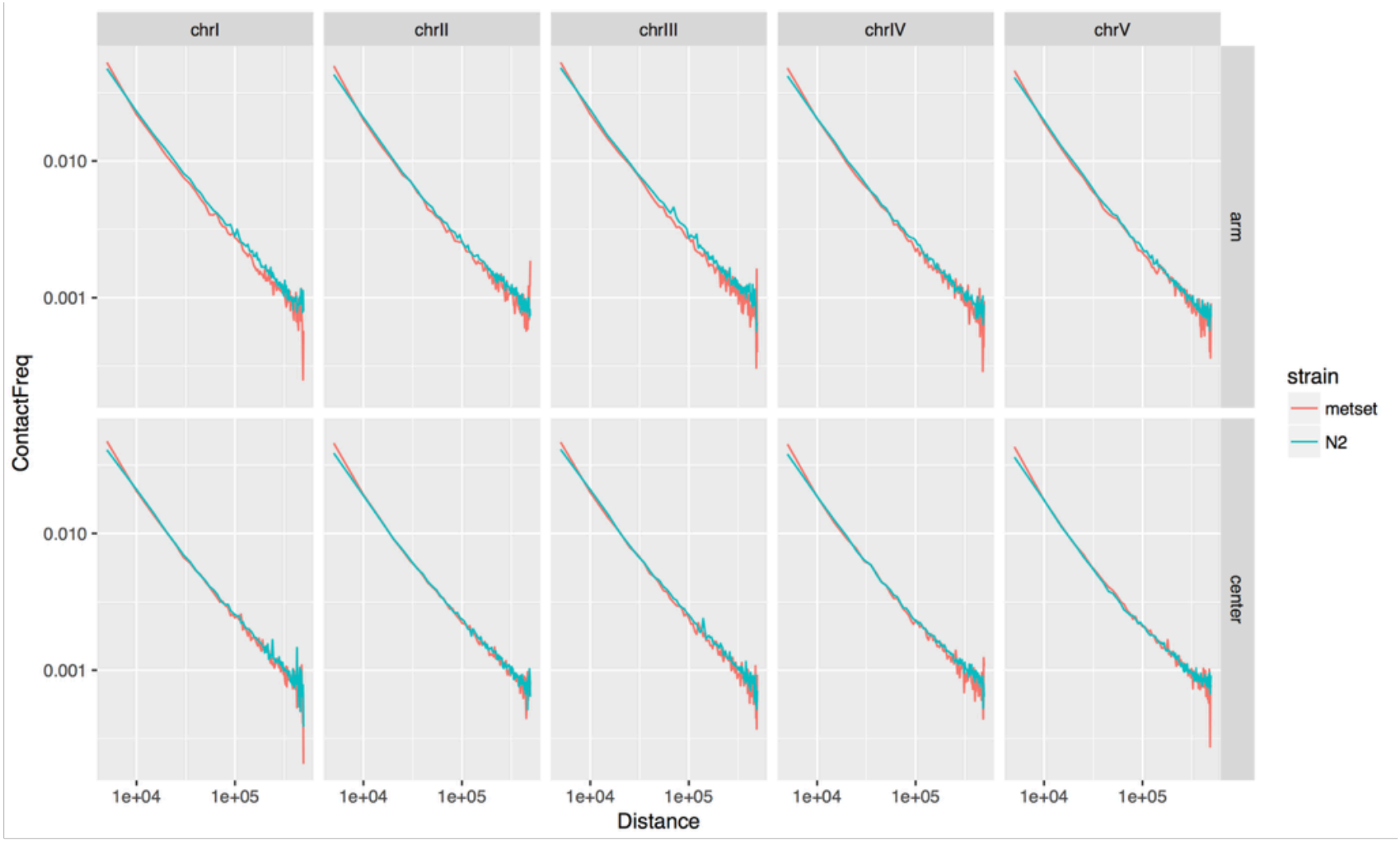


Fig 4.12: Contact probability as a function of genomic distance. Short range: 10kb-100kb. Contact frequency for wild-type (N2) and *met-2 set-25* (metset) were normalised to the same matrix column/row sum.

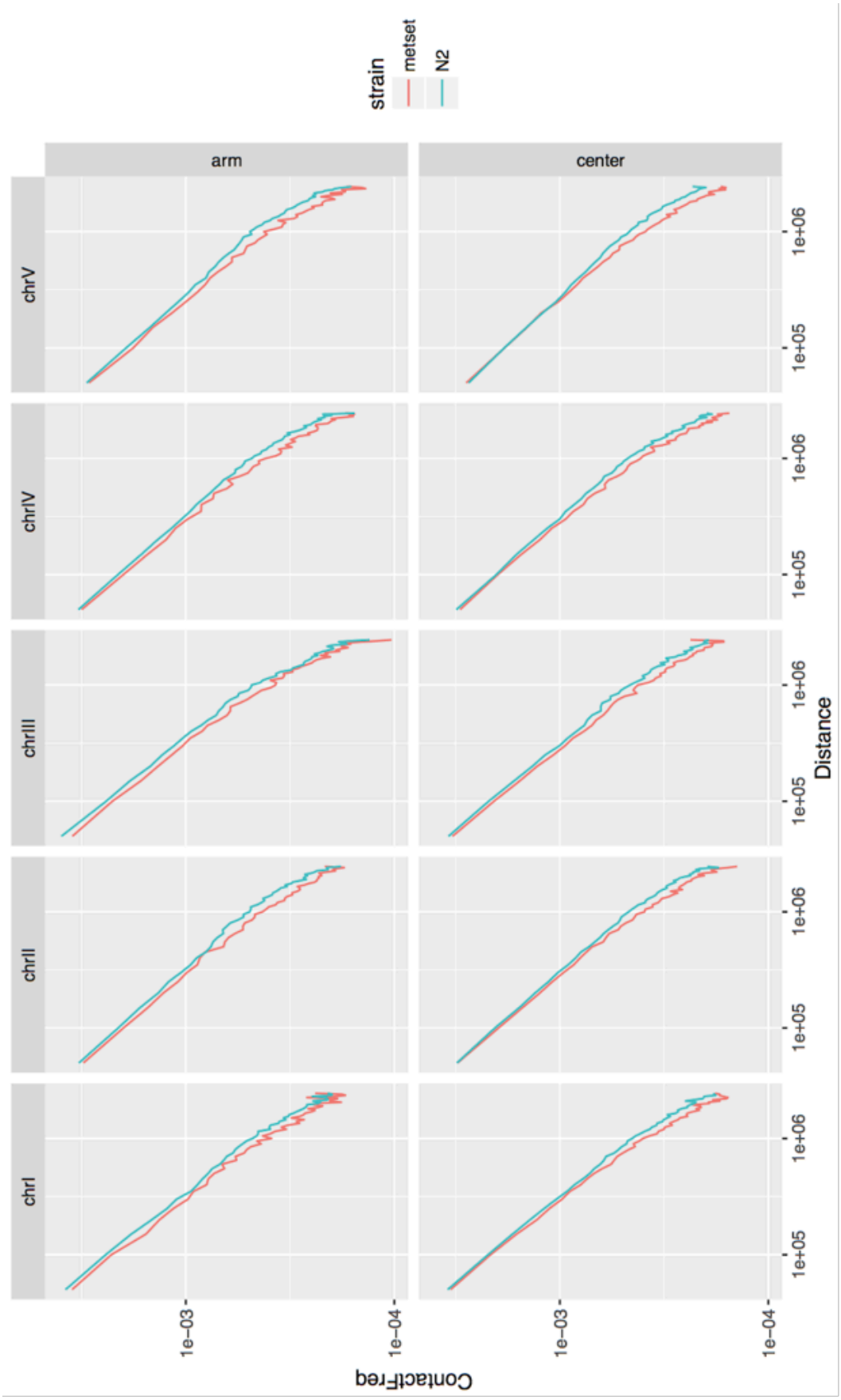


Fig 4.13: Contact probability as a function of genomic distance. Long range: 100kb-1Mb. Contact frequency for wild-type (N2) and *met-2 set-25* (metset) were normalised to the same matrix column/row sum.

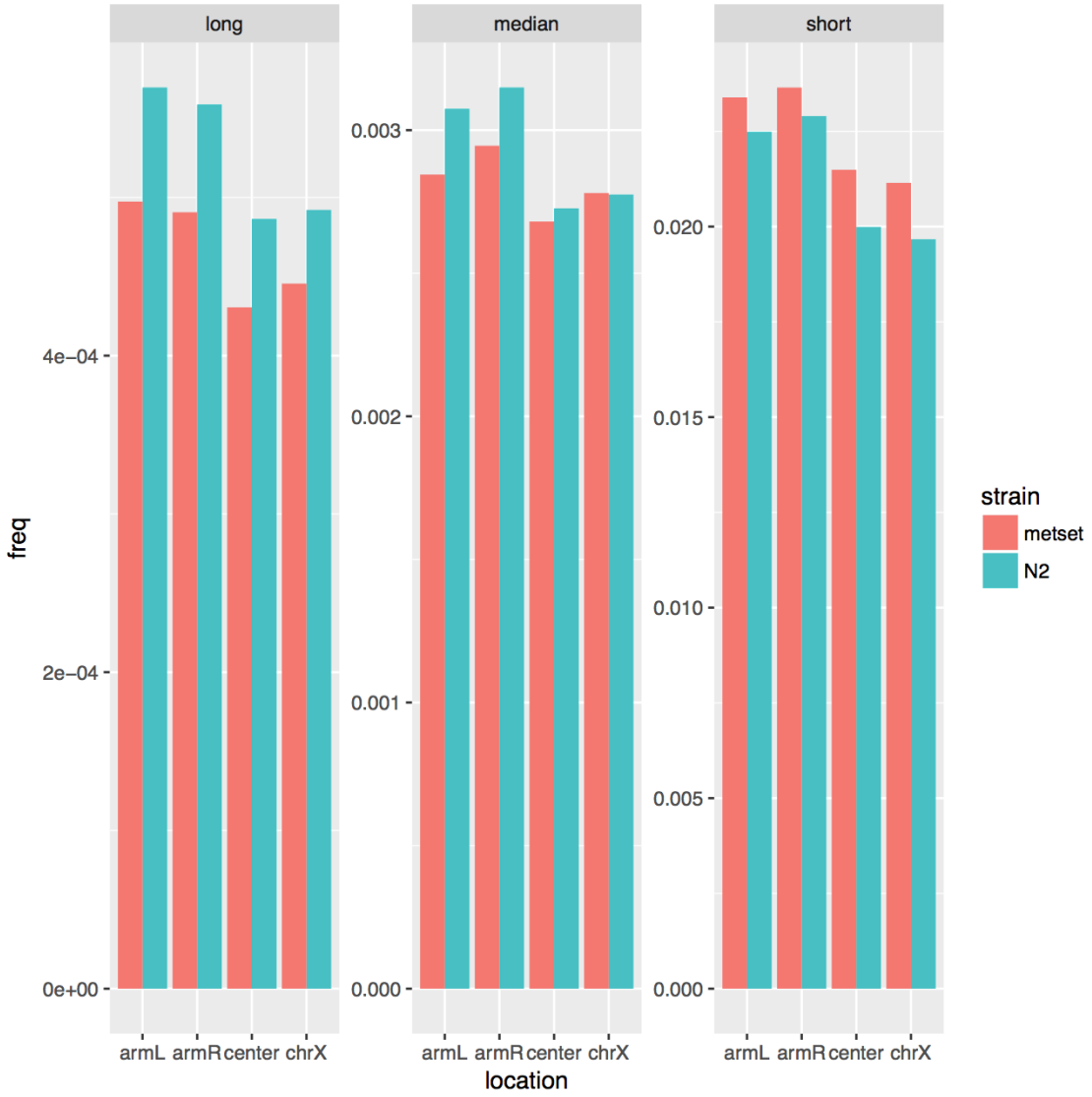


Fig 4.14: Frequency of interactions in *met-2 set-25* (metset) and wild-type (N2) at long (>200kb), median (20-200kb), and short (<20kb) ranges separated by location - left arm, right arm, centre, and chr X.

Fig 4.15: Domain and compartment APA (50kb-2Mb, 2-10Mb) for active and regulated chromatin state domains in wild-type (N2) and *met-2 set-25*. APAs are further separated into arms, centre, and overall (chrom). The numbers measure the percentage enrichment of the centre blocks over neighbourhoods.

The precise extent to which H3K9me affects interaction domains and compartments is unclear. Moving forward, we need to parse the individual contributions of H3K9me3 and H3K9me1/2 through ARC-C in *met-2* and *set-25* mutants separately. Moreover, there is a complex interplay between H3K9me and various histone modifications and chromatin regulators. For instance, although Towbin *et al* (2012) reported no change in H3K27me3 in *met-2 set-25* embryos, we observed a possible reduction in H3K27me3 ChIP-seq levels in regulated chromatin state domains (**Fig 4.16**), suggesting that H3K9me may have an indirect role. While we show here that H3K9me is involved in compartmentalisation,

further work needs to be done to elucidate the mechanism and role of other factors.

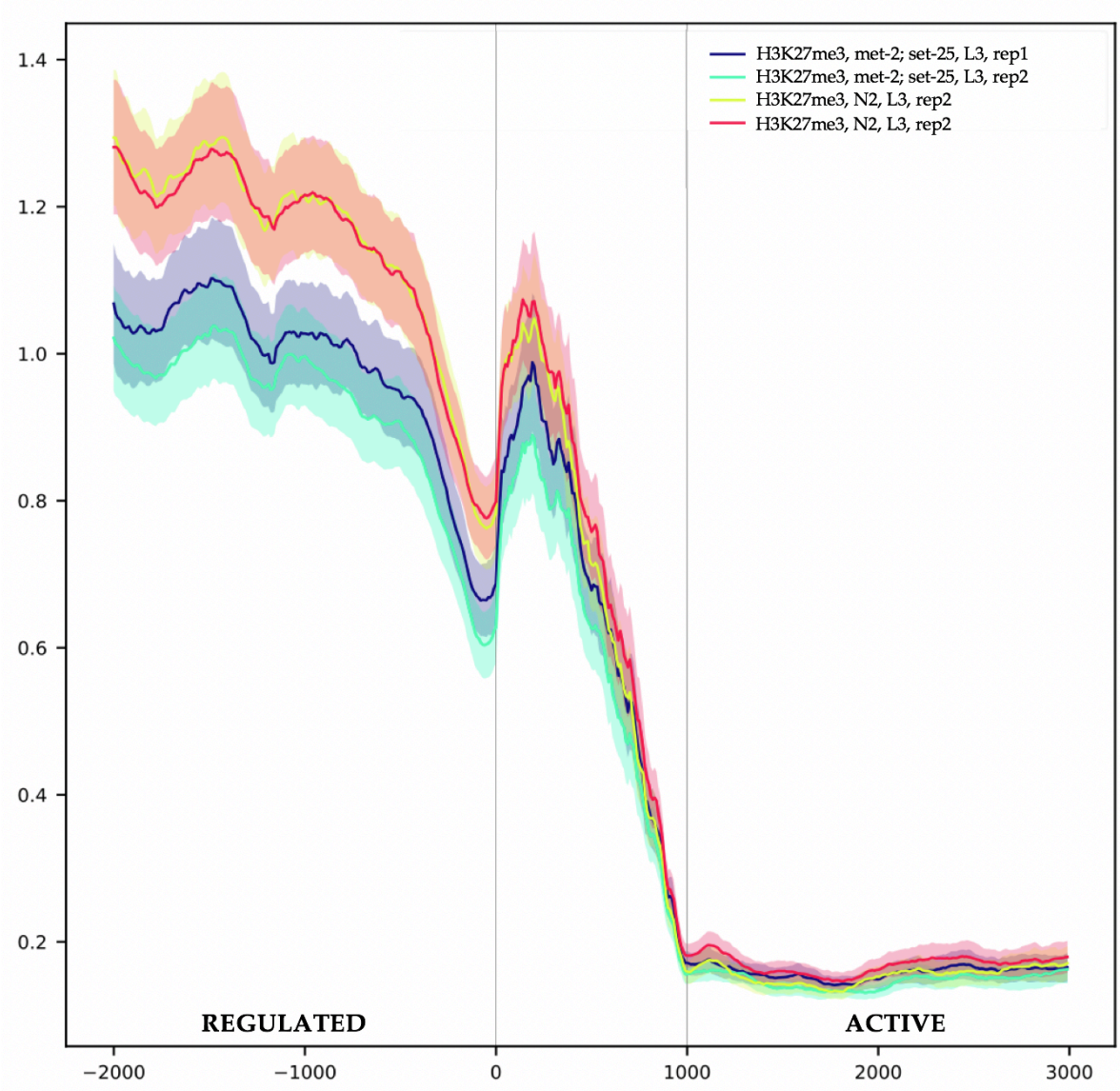


Fig 4.16: Coverage of H3K27me3 ChIP-seq over regulated, active chromatin state domains and borders in *met-2 set-25* and wild-type (N2).

CHAPTER V: NEXT-GENERATION ARC-C

In earlier chapters, I describe the regulatory landscape and chromatin architecture in L3 stage worms. Besides L3 stage larvae, I have applied ARC-C in embryos and starved L1 worms (data not shown). I have also performed ARC-C in *Drosophila S2* and human GM12878 cells (data not shown). In GM12878, while ARC-C enriches for regulatory interactions, it does not recover enough informative reads in its current form for high resolution analyses (data not shown), given the larger interaction space in mammalian genomes. ARC-C has to be improved upon if it were to be used in larger genomes.

Optimisation strategies

The greatest impediment to adoption is its relatively low efficiency in terms of informative reads. I adopted two main strategies. First, I attempted to increase the efficiency of the current steps by changing the metal cation in the digestion buffer or by adopting a different nuclease. Second, I redesigned the entire protocol and included a biotin-streptavidin pull-down step to enrich for informative reads in the manner of Hi-C. Out of these preliminary experiments, I found a method involving Tn5 transposase loaded with biotinylated nucleotides the most promising and performed further optimisation experiments.

For the former, I tried using Mn^{2+} instead of Mg^{2+} in the DNase I digestion buffer. DNase I has a divalent ion requirement (Junowicz *et al* 1973) and the type of cation used affects the properties of DNase I digestion: with Mg^{2+}, Ca^{2+}, or Zn^{2+}, DNase I makes single-strand nicks randomly in the phosphate backbone, creating fragments with variable overhangs; with transition metal cations such as Mn^{2+} or Co^{2+}, DNase I tends to generate double-stranded breaks with overhangs of 0-2 bp (Campbell & Jackson 1980).

I also tried an alternative nuclease - DNA fragmentation factor (DFF), which is activated during apoptosis to engender DNA fragmentation. DFF comprises DFF40/caspase-activated DNase (CAD) and its inhibitor DFF45/ICAD; re-engineered DFF can be activated by TEV protease cleavage of ICAD. The benefits of DFF are that they produce double-stranded DNA cuts with blunt ends that can be directly used in ligation and also cut between nucleosomes (Allan *et al* 2012); varying multiples of nucleosomes can be seen at the highest concentration (8 ul of 1U/ul DFF incubated for 30 min) of chromatin digestion (**Fig 5.1**).

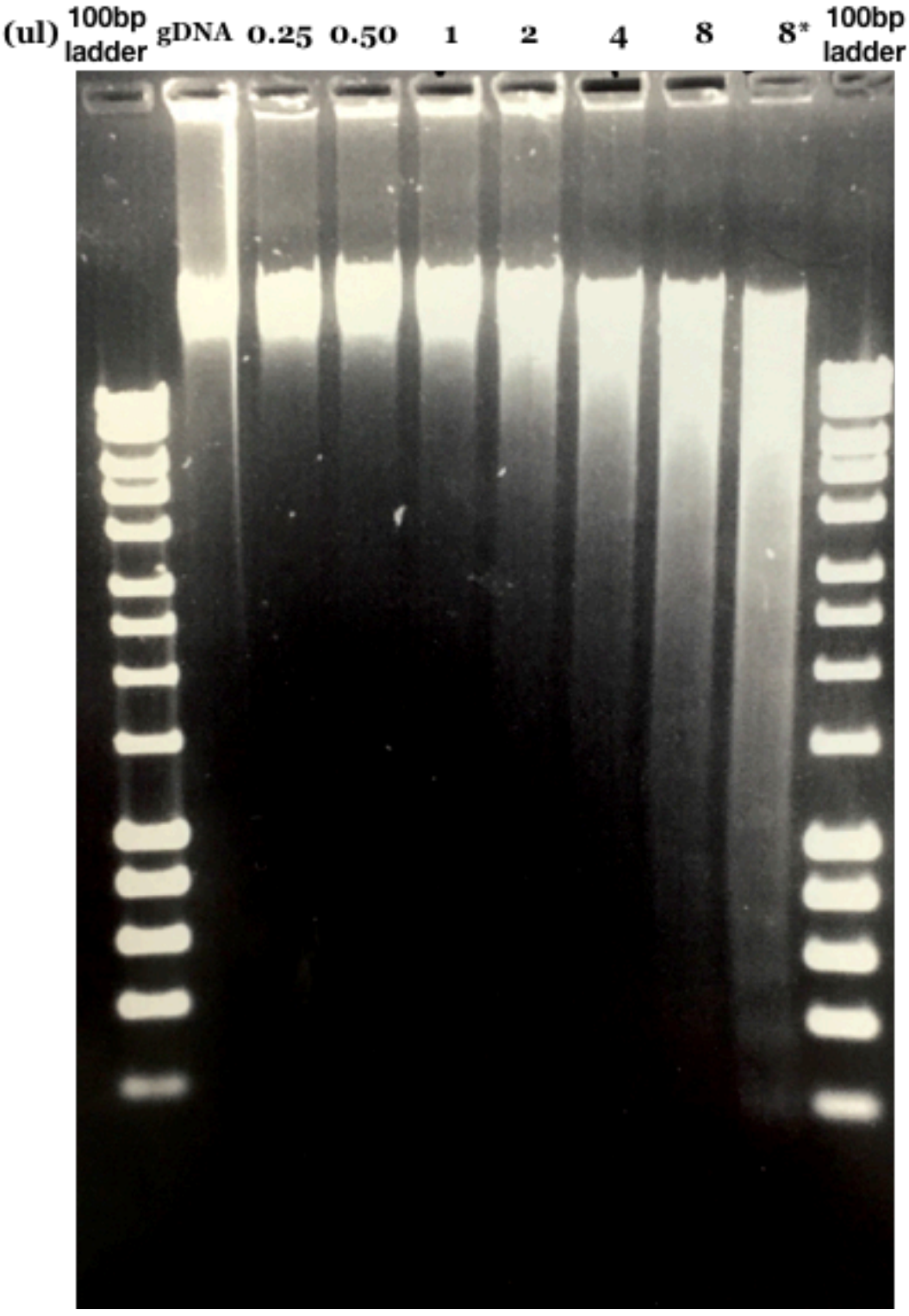


Figure 5.1: Gel electrophoresis of ~ 1ug DFF-digested chromatin on 1.5% agarose gel. "gDNA" indicates no DFF used; 0.25-8U of DFF were incubated with chromatin for 15min at RT; "8*" indicates 8U of DFF incubated with chromatin for 30min at RT

I tried four different means of incorporating biotin-tagged nucleotides into chromatin. First, I filled in overhangs from DNase I digestion with biotinylated nucleotides. Since DNase I digests nonspecifically, I decided to use both biotin-11-dATP and biotin-11-dCTP to capture more loci (although there might be a slight under-representation in A and C-only regions; it is difficult to predict where DNase I will cut - it might not produce complementary overhangs since it can digest DNA down to mono-, di-, and tri-nucleotides (Kunitz *et al* 1950) and may on occasion produce blunt ends) and also because they were cheaper and more readily available for purchase. I did not use more biotinylated nucleotides (i.e. biotin-dTTP and biotin-dGTP) as excessive incorporation could create steric interference between biotin molecules and affect the efficiency of downstream processes. Subsequently, I found that this condition was similar to Micro-C, which uses MNase and biotinylated nucleotides (Hsieh *et al* 2015). Second, I experimented with biotin-14-dATP and biotin-14-dCTP. These hapten-tagged nucleotides had longer linkers (14-atom), which could improve the efficiency of incorporation, blunting, and ligation.

Third, to avoid the inconsistency and unpredictability of using biotinylated nucleotides to fill in overhangs, I tried using biotinylated oligonucleotides (GCTGAGGGAT^{b}C) instead as a bridging adaptor between interacting fragments

(A-tailing was performed as well to reduce self-ligation or ligation without the adaptor present) and this was adapted with modifications from an earlier DNase-Hi-C paper (Ma *et al* 2015).

Lastly, I devised a new method that uses the Tn5 transposase (from the Nextera Mate Pair Library Preparation Kit), which are fitted with biotinylated oligonucleotides, instead of DNase I; this new method is called next-generation ARC-C (ngARC-C) (**Fig 5.2**). Results for all optimisation experiments with the exception of ngARC-C are summarised in **Table 5.4**. Results for ngARC-C optimisation experiments are summarised in **Table 5.5**.

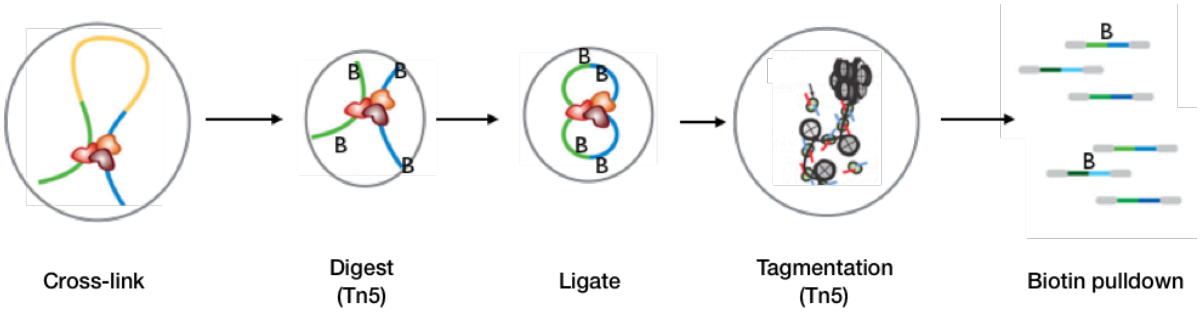


Figure 5.2: Graphical schematic of ngARC-C protocol.

Next-generation ARC-C (ngARC-C)

Similar to ARC-C, chromatin is first fixed with formaldehyde. Instead of DNase I, Tn5 transposase is used to insert biotinylated nucleotides into open

regions of chromatin, which can then be proximity ligated and subsequently pulled down with streptavidin-coated beads to enrich for informative DNA fragments.

Tn5 transposase plays a central role in ngARC-C. Wild-type Tn5 is composite transposon of two insertion sequence 50 (*IS50*) elements, which are themselves part of the *IS4* family of transposable elements (Naumann & Reznikoff 2000, Chandler & Mahillon 2002). Transposition happens through a 'cut and paste' mechanism which comprises nonspecific DNA binding/end sequence (ES) recognition, synapsis, cleavage, target DNA capture, and strand transfer, resulting in a 9bp micro-duplication of the target sequence (Reznikoff 2008). Due to steric effects, transposition events require a minimum spacing of around 38bp (Adey *et al* 2010). However, wild-type Tn5 has very low transposition activity. This is partially alleviated by several modifications: E54K mutation to improve transposase recognition of ES (Zhou & Reznikoff 1997), L372P mutation to separate the inhibitory interaction between the N- and C- termini (Weinreich *et al* 1994), and an optimal mosaic ES (ME) of the natural inner and outer end sequences (Zhou *et al* 1998) for improved transposase dimerisation. This hyperactive form of Tn5 transposase can be loaded with oligonucleotides containing the ME and is used for tagmentation reactions.

The main issue with using hyperactive Tn5 is that it holds onto DNA at two positions after tagmentation until an additional step is undertaken to strip it off. Typically, this is achieved via at least 0.1% of SDS (Picelli *et al* 2014) or heat inactivation at 65-72C (Naumann & Reznikoff 2000, Picelli *et al* 2014). As shown in single molecule imaging of Tn5 post-transposition (**Fig 5.3**), in the absence of SDS, DNA fragments appear as long strands (held together by bound Tn5) until they are treated with SDS, after which they become smaller fragments (Amini *et al* 2014). In the case of ATAC-seq, treatment with a chaotropic reagent (Qiagen buffer QB, which probably contains guanidine hydrochloride and isopropanol) is also sufficient to release Tn5, but presumably also destroy nuclear structures as well.

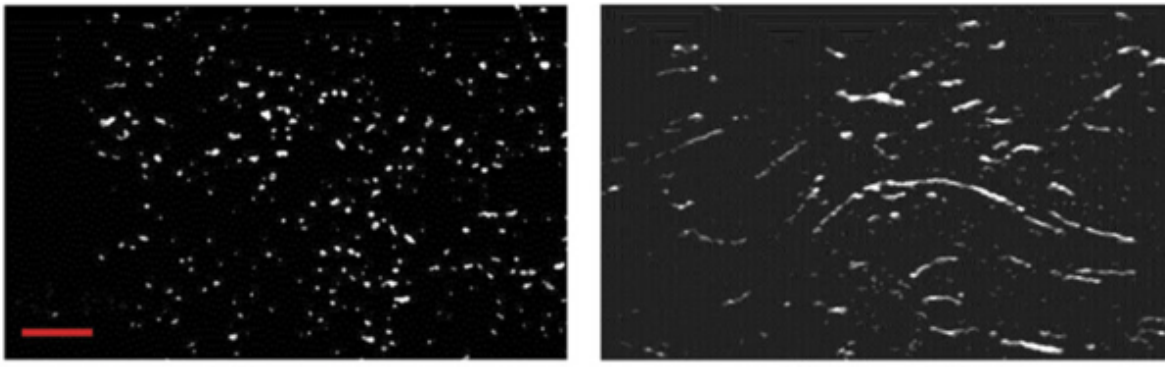


Figure 5.3: Single-molecule imaging of YOYO-1 fluorescent dye labelled DNA after Tn5 transposition. Tagmented DNA was treated with SDS to a final concentration of 0.05%. Scale bar 10um. (Amini *et al* 2014).

I experimented with either the use of 0.1% SDS or heat activation at 72C to strip Tn5 off DNA, and also whether to use Tn5 (from Nextera Library Prep Kit) in the manner of ARC-C or to directly sonicate the nuclei after ligation and make ngARC-C via the Truseq Library Preparation approach. The Truseq approach was driven by the concern that the use of Tn5 would be inefficient and reduce complexity, given the steric effects of biotin tags and that only half of tagmentation products are sequence-able. In Nextera kits, two different tags that eventually act as PCR primers are randomly inserted into DNA; fragments that contain two different tags can be sequenced but those containing two of the same tags cannot, effectively losing half of the input DNA.

Observations

I made and sequenced ARC-C libraries with these different methods and variations. To evaluate them, I looked primarily at their efficiency, as defined by the number of informative over valid reads (considering other statistics such as mitochondrial reads, or mappability would confound the evaluation by assimilating factors that are likely unrelated to the modifications), and the aggregate coverage over DHS. I note, however, that these variations were attempted only once, with the exception of ngARC-C, and so, observations from these series of experiments may not be statistically robust.

Method	Enrichment over DHS	*Cis* (%)	Efficiency = Info / Valid (%)
ARC-C (Mg^{2+}) - N2_2a	3.9	78.07	8.669
Mn^{2+} (with end repair)	2.7	68.06	7.887
DFF with end repair	2.9	49.55	4.959
DFF without end repair	3.1	32.97	8.502
biotin-11	2.0	73.63	14.894
biotin-14	3.5	71.13	19.067
biotin-adaptor	2.9	60.57	10.559

Table 5.4: Key statistics of optimisation experiments, excluding ngARC-C. Enrichment over DHS measures the enrichment of informative read coverage at DHS over background, which reflects signal to noise. Cis (%) measures the percentage of *cis* informative reads over all informative reads. Efficiency measures the number of informative over valid reads and approximates the proportion of informative chimeric DNA within the library.

Taking the wild-type L3 stage libraries as the baseline reference (median efficiency = 8.126%, $n = 6$, $sd = 0.937$), I found that the efficiency for blunting was likely optimal. The use of Mn^{2+} (7.887%) and DFF nuclease (with end repair = 4.959%, without end repair = 8.502%) do not appear to perform much better than DNase I with Mg^{2+} (**Table 5.4**). The library with end repair could have performed worse because end repair is an equilibrium process between 3′ to 5′ exonuclease activity, which is more active on single than double stranded DNA, and 5′ overhang fill-in (Wang *et al* 1994). The use of end repair enzymes could

inadvertently create DNA with 5' overhangs that cannot be ligated. That DFF nuclease not have a higher efficiency with an additional end repair step as control provides evidence that it does produce ligation-ready blunt ends. In all, the results suggest that it would plausibly be futile to pursue this avenue further.

The biotin-streptavidin strategy increased efficiency in all scenarios. As expected, with the use of biotin tags with a longer linker to reduce steric hindrance, the efficiency when using biotin-14 (19.067%) was higher than when using biotin-14 (14.894%); the aggregate coverage over DHS for biotin-14 was higher as well (3.5 vs 2.0) (**Table 5.4**). Regrettably, the use of biotinylated adaptors did not improve efficiency by much (10.559%) (**Table 5.4**). It is possible that either the A-tailing efficiency was low or the ligation process was inefficient.

Excitingly, the approach of using hyperactive Tn5 transposase with biotinylated oligonucleotides (ngARC-C) resulted in the greatest increase in efficiency (67.180% with heat inactivation) (**Table 5.5**). Whilst the biotinylated nucleotide fill-in or bridge adaptor approaches were encouraging, ngARC-C showed the most promise. Indeed, all Nextera-made ngARC-C libraries with biotin pull-down and some form of inactivation have a complexity lower than 25.5% (or a duplication rate of 74.5%) (**Table 5.5**). One Truseq-made library with the same conditions had a complexity of 75.97%, albeit with only approximately

400,000 valid reads. Complexity can, however, be increased to an extent by the use of more Tn5 transposase (MP) for fragmentation: complexity is lower when using 0.5 ul of Tn5 (MP) (18.67% with 144,026 valid reads) as compared to 2.5 ul (30.24% with 654,648 valid reads) (**Table 5.5**).

ID	Tn5 (MP) (ul)	Inactivation		Biotin Pulldown	Library Prep Method	Efficiency (%) = Info/Valid	# Valid reads	Complexity (%)	Enrichment over DHS
		72C	0.1% SDS						
A	2.5	-	-	+	Nextera	11.247	1,813,282	13.71	1.5
B	2.5	+	-	+	Nextera	67.180	1,118,412	22.50	2.0
C	2.5	+	-	-	Nextera	2.491	17,128,934	82.55	2.1
D	2.5	+	-	+	Truseq	40.912	398,566	75.97	2.7
E	2.5	-	+	-	Truseq	1.624	15,310,150	93.57	1.5
F	2.5	-	+	-	Nextera	1.088	102,081,770	94.13	1.4
G	2.5	-	+	+	Nextera	8.221	7,834,542	25.50	1.6
H	2.5	-	+	+	Truseq	38.027	654,648	30.24	1.3
I	0.5	-	+	+	Truseq	29.274	144,026	18.67	1.5

Table 5.5: Summary of ngARC-C optimisation experiments with key statistics. Table indicates (left to right): ID for libraries (**Appendix**), amount of Tn5 transposase used, method of inactivation - 72C heat inactivation or 0.1% SDS, whether streptavidin pulldown was performed, and the type of library preparation performed. Efficiency measures the number of informative over valid reads. Complexity measures the amount of unique read pairs after removal of PCR duplicates. Enrichment over DHS measures the enrichment of informative read coverage at DHS over background.

The hypothesis that the assimilation of biotinylated nucleotides induces steric limitations is further buttressed by the much lower efficiency in ngARC-C libraries without biotin pull-down (C, E, F: 1.088 - 2.491%) (**Table 5.5**), which should at least have been comparable with classical ARC-C libraries in theory (median efficiency for wild-type libraries = 8.126%, **Appendix - Table A1.1**). But, as expected, biotin pull-down increased efficiency by 7.556 to 26.969 fold (**Table 5.5**) when controlled for all other variables. Finally, there is a need to proactively inactivate Tn5 - a ngARC-C library without any inactivation steps gave an efficiency of 11.247% while other libraries that do have inactivation have efficiencies ranging from 8.221 to 67.180% (median = 38.027%) (**Table 5.5**). However, this inactivation - with current conditions, at least - appears to come at the expense of nuclear architecture: aggregate signal over DHS in all samples ranges from 1.3 to 2.7 (**Table 5.5**), lower than acceptable (>3 based on ATAC-seq) and also consistent with the series of optimisation experiments in *Drosophila* S2 and GM12878. With some additional work, ngARC-C has the potential to be much more efficacious than the current version of ARC-C, while keeping the fundamental principles intact.

In vitro ngARC-C optimisation

Without making more ngARC-C libraries, I decided to identify potential areas of improvement (**Fig 5.6**) and ran *in vitro* optimisation experiments. To start with, it is important to ensure that Tn5 transposase can insert biotinylated oligonucleotides into open chromatin and that the lower aggregate coverage over DHS is not a product of using biotinylated oligonucleotides instead of sequencing adaptors. And indeed, when I tagmented 500,000 *C. elegans* N2 L3 stage nuclei with 1 ul Tn5 transposase (MP), purified and size-selected for 100 to 300 bp DNA fragments, and ran a qPCR to measure the signal over a diagnostic DHS, the results were reasonable (6.32; 4 to 7 denotes good quality ARC-C libraries).

As mentioned, the inactivation of Tn5 is crucial to the success of the protocol (**Fig 5.6**). Next, unligated yet biotinylated DNA ends have to be removed as they provide little useful information about 3D conformation but can be pulled down by streptavidin beads. Streptavidin beads can also be blocked to reduce non-specific binding of non-biotinylated DNA fragments. Lastly, I found that human ARC-C libraries are replete with mitochondrial DNA (data not shown). The CRISPR-Cas9 system can be used to target the mitochondrial genome to make ARC-C in humans feasible.

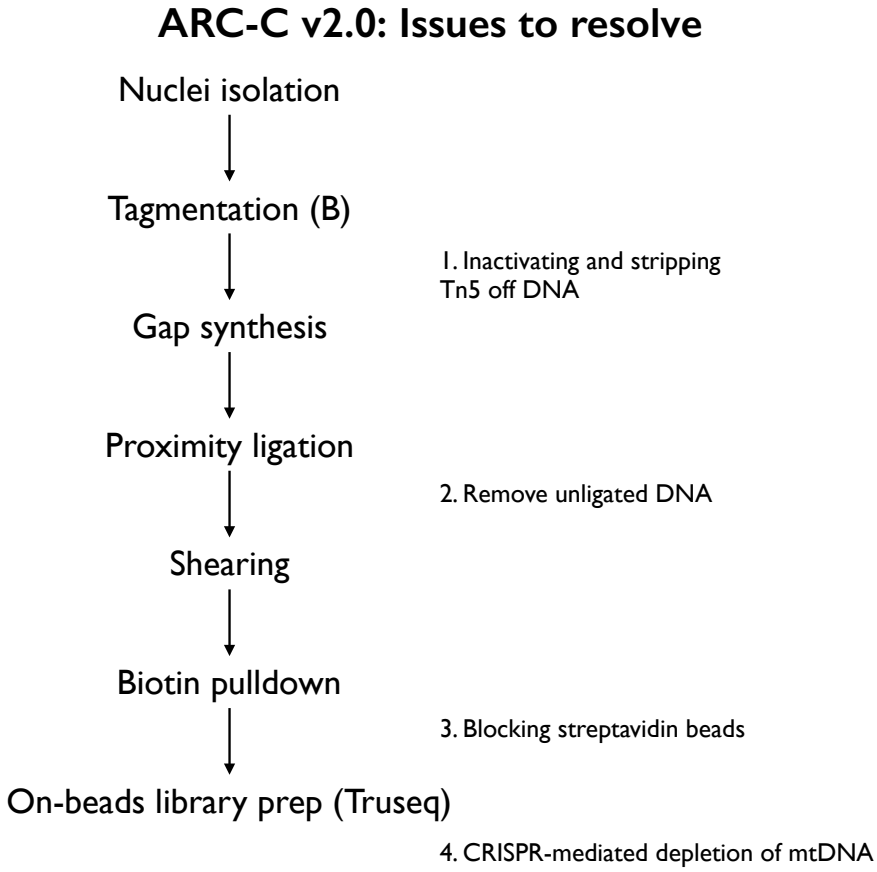


Figure 5.6: Schematic of ngARC-C protocol and issues to resolve at corresponding steps.

The electrophoretic mobility shift assay can be used to test the conditions necessary for Tn5 inactivation. I incubated 0.1 ul of Tn5 transposase (Nextera) with 100 ng of *C. elegans* wild-type genomic DNA (gDNA) for 10 min at 37C before trying the different inactivation conditions and running the products on a 1.5% agarose gel. In the absence of inactivation, Tn5 transposase was bound to gDNA, causing a band shift (**Fig 5.7:** 2). When Tn5 transposase is stripped with 0.1% SDS (incubated for 5 min at 25C) (**Fig 5.7**: 3) or at 72C (**Fig 5.7:** 4), tagmented

DNA fragments were released. I tested milder detergents - 1% CHAPS (**Fig 5.7**: 5; **Table**) or 0.025% SDS (**Fig 5.7**: 6) - and 150 mM EDTA (**Fig 5.7:** 7) to chelate Mg^{2+} ions, as they are required for Tn5 transposase activity (Goryshin & Reznikoff 1998). Promisingly, they were all able to strip Tn5 transposase off DNA.

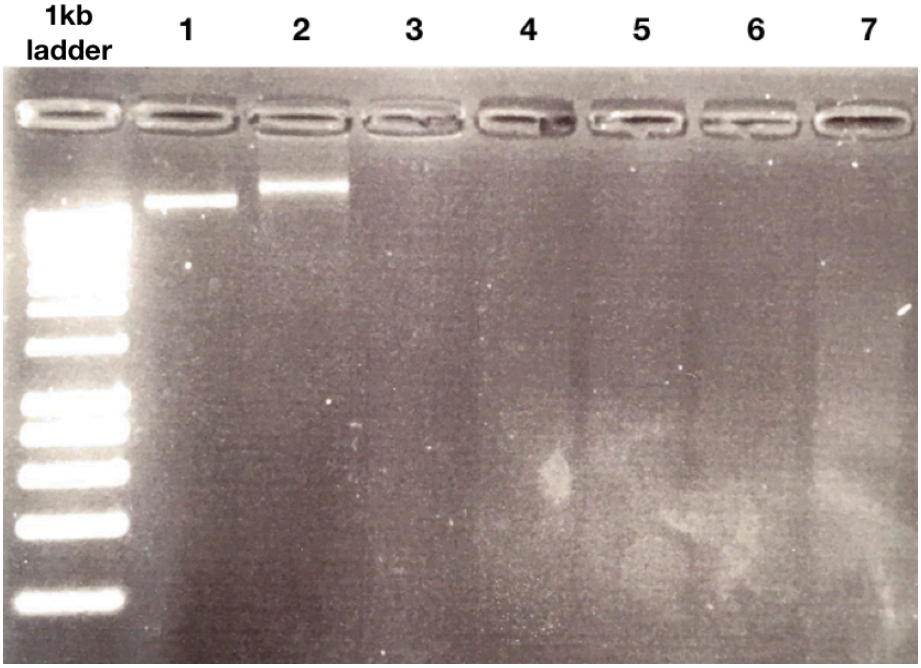


Figure 5.7: Gel electrophoresis of 100ng of DNA after Tn5 transposition on 1.5% agarose gel. (1) No Tn5 added. (2) Tn5 added, no inactivation. (3) 0.1% SDS. (4) 72C. (5) 1% CHAPS. (6) 0.025% SDS. (7) 150mM EDTA.

I devised an assay to evaluate the efficacy of pre-blocking streptavidin beads on the reduction of non-specific DNA binding. I first obtained an equimolar mixture of two unique 200 bp oligonucleotides (one with a biotin tag on each end and one with no biotin modifications). Streptavidin beads were then pre-blocked

with 0.2 mg/ml salmon sperm DNA before pull-down was performed with the oligonucleotides. I extracted DNA with 95% formamide, purified them with AMPure XP beads and ran qPCR to quantify the abundance of each oligonucleotides with primers that were controlled for the same amplification efficiency. Without pre-blocking, the fraction of biotinylated to non-biotinylated oligonucleotides were 4.12; with pre-blocking, the ratio went up to 21.57 (**Fig 5.8**), a fold change of 5.24. In the context of ngARC-C, pre-blocking with salmon sperm DNA would be applicable if DNA was A-tailed prior to biotin pull-down to prevent contamination. Alternatively, streptavidin beads can be pre-blocked with yeast RNA, although it may be less efficient at reducing non-specific DNA binding.

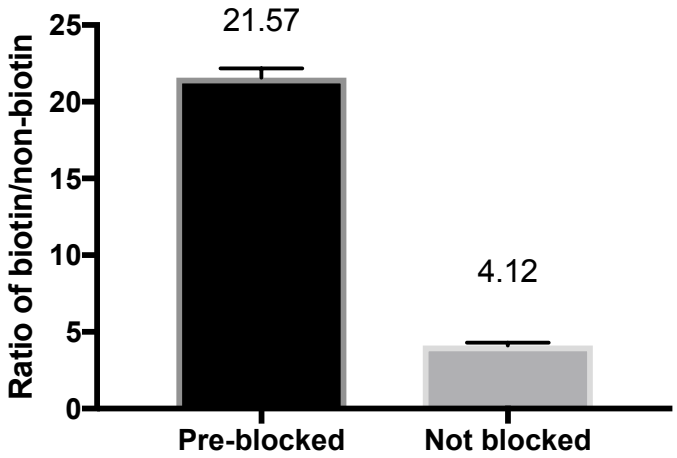


Figure 5.8: Ratio of biotinylated to non-biotinylated oligonucleotides after pulldown, with or without pre-blocking.

The CRISPR-Cas9 tool can be manipulated to target specific DNA sequences and degrade them without editing. This technique, called CRISPR/Cas9-assisted mitochondrial DNA depletion (CARM), has been successfully applied in mouse chromatin studies (Wu *et al* 2016). A library of 395 single guide RNAs (sgRNA) was designed to target and cover around every 40 bp of the mitochondrial genome. Genomic DNA that was extracted from whole GM12878 cells were incubated with 500 ng sgRNAs and 1 ug of Cas9 protein for 2 h at 37C before the reaction was quenched with a stop buffer (30% glycerol, 1.2% SDS, 250 mM EDTA); DNA was later purified with 1.2x AMPure XP beads. I then ran qPCR to quantify the relative abundance of mtDNA relative to nuclear DNA using validated primers from Thakar *et al (*2015), which were meant for a similar analysis of mtDNA content. CARM resulted in a 327.47 fold reduction in the relative abundance of mtDNA post-depletion (**Fig 5.9**).

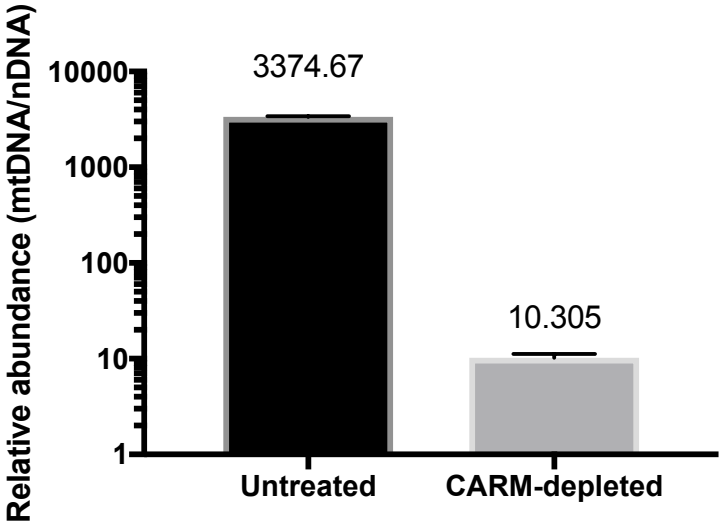


Figure 5.9: Relative abundance of mtDNA to nuclear DNA in untreated and CARM-depleted conditions.

Discussion

To increase the recovery of informative reads, I performed a series of optimisation experiments for ARC-C. Attempts to improve the efficiency of end repair were unsuccessful as the process was likely optimal. Taking a leaf from Hi-C and other related methods, I pursued the use of biotin-streptavidin pull-down to enrich for informative DNA fragments (i.e. non-contiguous fragments that arose as a result of digestion, end repair, and ligation). To this end, I have tried to fill in overhangs with biotinylated nucleotides, the use of biotinylated bridge adaptors, or a revamped protocol that substitutes biotinylated oligonucleotides-carrying Tn5 transposase for DNase I (ngARC-C). Of these, ngARC-C proved to be the most promising, with the largest increase in the efficiency of capturing informative reads.

Conceptually, ngARC-C would be worth pursuing because it is simpler to perform and is presumably more robust to day-to-day variations, unlike the use of DNase I in ARC-C, which has to be carefully calibrated to ensure a optimal level of digestion. The direct insertion of biotinylated oligonucleotides means that fewer steps and parameters have to be considered. However, there are still several kinks to iron out. The aggregate coverage over DHS for ngARC-C libraries are relatively low, likely due to the harsh conditions I used initially to strip Tn5 off DNA. Gentler conditions would ameliorate this problem. As I have alluded to

earlier, the use of ARC-C (or Tn5 transposase, in general) in fixed cells produce substantial mitochondrial contamination, but this is a known problem and several groups have strategies to tackle it, one of which includes the use of Cas9 to target mtDNA, which I have successfully applied *in vitro*. Due to practical constraints, I did not manage to formalise a protocol for ngARC-C, but I believe that the results here should provide a good foothold to complete ngARC-C.

DISCUSSION

In this thesis, I describe the development of a novel adaptation of Hi-C - the Accessible Region Conformation-Capture - that enriches for regulatory interactions genome-wide. Significant interactions were called in ARC-C with a method that takes into account the distance-dependent decay of contact frequency, normalises for accessibility, and estimates background by looking at interactions outside of ATAC-seq peaks (**Chapter I**). 98.21% of these were between annotated regulatory elements (**Chapter II**). By utilising ARC-C's increased sensitivity at regulatory elements, we screened for factors that could putatively be involved in loop formation with the use of aggregate analyses and identified known looping factors (e.g. CTCF, cohesin) as well as other novel ones (**Chapter III**). Although ARC-C enriches for coverage over regulatory elements, it can be used to study domains and compartments. We found evidence that active and regulated chromatin state domains form insulated contact domains and compartments (**Chapter IV**). Currently, ARC-C is limited by its ability to recover informative reads and large inroads have been made toward a more efficient protocol (**Chapter V**).

Whitherto, ARC-C?

The understanding of 3D genome organisation in various organisms has been propelled by advances in microscopy and conformation capture techniques. The choice of technique will ultimately come down to the research question. Hi-C is useful if one intends to study large scale structures and model chromatin folding; for instance, it has allowed us a better understanding of mitotic chromosome organisation (Naumova *et al* 2013, Nagano *et al* 2017, Gibcus *et al* 2018). To study the regulatory interactome, an enrichment step is necessary. ARC-C theoretically allows for single bp resolution and thus very fine mapping of regulatory interactions. Moreover, while ARC-C asks a broader question by not pre-determining specific regions of the genome to interrogate, it can eventually be combined with probe-based pulldown methods (ie. Capture-C, Capture Hi-C) for even higher resolution at selected loci of interest.

Moving forward, the next-generation ARC-C (**Chapter V**) should be pursued. It is in many ways more advanced than ARC-C. The use of Tn5 transposase for fragmentation and biotinylated adaptor ligation would abrogate the concern over digestion levels and reduce sample-to-sample variability. In addition, the sensitivity of Tn5 transposase, presumably due to its ability to simultaneously fragment and ligate, may allow it to be applied to as few as 500 cells as in Omni-ATAC (Corces *et al* 2017) as opposed to conventional DNase-seq,

which requires millions of cells (Jin *et al* 2015). Once optimised, implementation should be quick (1-2 days) given the simplicity of the protocol and the increased efficiency should make it amenable to larger genomes. The low cell requirement and simple protocol would make it useful for interrogating regulatory interactions in clinical samples, especially when analysed in conjunction with GWAS.

Genome organisation in *C. elegans*

We found that regulated domains and active domains formed compartments. For regulated domains, the depletion of H3K9 methylation weakened regulated compartments (**Fig 4.15**). It is plausible to connect the loss of regulated compartments to the detachment of H3K9me3 domains from the nuclear lamina in *met-2 set-25* mutants (Towbin *et al* 2012). However, Falk *et al* (2018) found that compartments were stable in rod photoreceptors with inverted nuclei that do not have heterochromatic tethering to the nuclear lamina. In inverted nuclei, heterochromatin forms a dense core with euchromatin forming an outer ring, suggesting that compartments are not a product of nuclear localisation but likely an intrinsic property of chromatin. More likely, the reduced binding of heterochromatin proteins such as HPL-2, the *C. elegans* heterochromatin protein 1 (HP1) in *met-2 set-25* mutants (Garrigues *et al* 2015), led to weaker interactions between regulated domains. Human and *Drosophila* HP1α were implicated in

phase separation of heterochromatin (Larson *et al* 2017, Strom *et al* 2017). In worms, this can be assessed and verified in *set-25* mutants that lack H3K9me3 but retain substantial levels of heterochromatic nuclear periphery association (Towbin *et al* 2012).

H3K36me3-enriched active and H3K27me3-enriched regulated domains appear to be antagonistic. Weaker regulated compartments were concomitant with stronger active compartments (**Fig 4.15**). Domains could provide weak insulation or impose physical constraints against self-association of their flanking domains. Indeed, H3K36me3 and H3K27me3 are antagonistic such that H3K27me3 spreads into germline-expressed genes lacking H3K36me3 in *mes-4* mutants (Gaydos *et al* 2012). It would be interesting to test if active compartments are weakened and regulated compartments are strengthened in *mes-4* or *met-1* mutants that lack transcriptionally associated H3K36me3.

HOT regions seem to be an important feature in *C. elegans*. They are, by definition, bound by an unusually large number of factors (> 29 as defined in this thesis), enriched at interaction hubs (95.3%) (**Chapter II**), and at active domains ($p < 0.0001$). In humans, HOT regions were frequently found in long-range interactions and had a strong affinity for other HOT regions (Heidari *et al* 2014). HOT regions appear analogous to transcription factories, which were described as

transcriptional "hotspots" enriched in RNA polymerase II that form central nodes in complex networks (Larkin *et al* 2013). I observed that gene pairs at HOT-HOT interactions were not more correlated than by chance (**Fig 2.3**), suggesting that HOT regions are structural. This is partly supported by the lack of transcriptional consequence of hub deletions. Indeed, transcription factories persisted even in the absence of transcription (Mitchell & Fraser 2008).

There is some evidence that HOT regions may nucleate chromatin activity. At the X chromosomes, SDC-2 might bind *rex* sites and open chromatin for initial condensin-I-like DCC recruitment (Albritton *et al* 2017). Strong *rex* sites frequently overlap HOT regions (Crane *et al* 2014). On the autosomes, I find HOT regions enriched at KLE-2 (condensin II) and SCC-1 (cohesin) binding sites (**Chapter III**). Given the similarities between the genetic elements and condensin complexes, it is plausible that a recruitment factor opens chromatin at HOT regions in the autosomes. Indeed, we found the motif for EOR-1 at HOT regions (data not shown). EOR-1 plays a role in chromatin accessibility, possibly as a pioneer factor as proposed in Daugherty *et al* (2017), and genetically interacts with nucleosome remodellers (Lehner *et al* 2006). Moreover, cohesin is known to recruit transcription machinery, including RNA polymerase II (Heidari *et al* 2014). Such a hypothesis can be tested in *eor-1* or *kle-2* knockout mutants, with the expectation that hubs would be lost.

The extent to which hubs, and by extension HOT regions, regulate transcription is unclear. Given the promiscuity of HOT interactions (**Fig 2.3**) and redundancy we observe for hub02 and hub03 (**Fig 2.17** & **Fig 2.19**), it is less likely for a single hub to be a key contributor to local transcription. In hub05 mutants, genes that were uncoupled to hub05 could retain access to transcription machinery in the context of transcription factories. Indeed, in mammals, the depletion of cohesin engendered extensive changes in chromatin architecture - loss of TADs but stronger compartmentalisation - with little effect on transcription (Schwarzer *et al* 2017, Rao *et al* 2017). Future work could involve sequentially deleting hubs (hub02 and hub03 would be good candidates) and tracking changes in transcription and chromatin architecture.

My work with ARC-C has established references from which other hypotheses can be tested. Open questions remain about whether loop extrusion occurs in *C. elegans* (a model for dosage compensation proposes that DCC load and spread through looping (Albritton *et al* 2017)), if the different cohesin subunits perform different function in organising the genome, and how the regulatory landscape in tissues appear.

METHODS

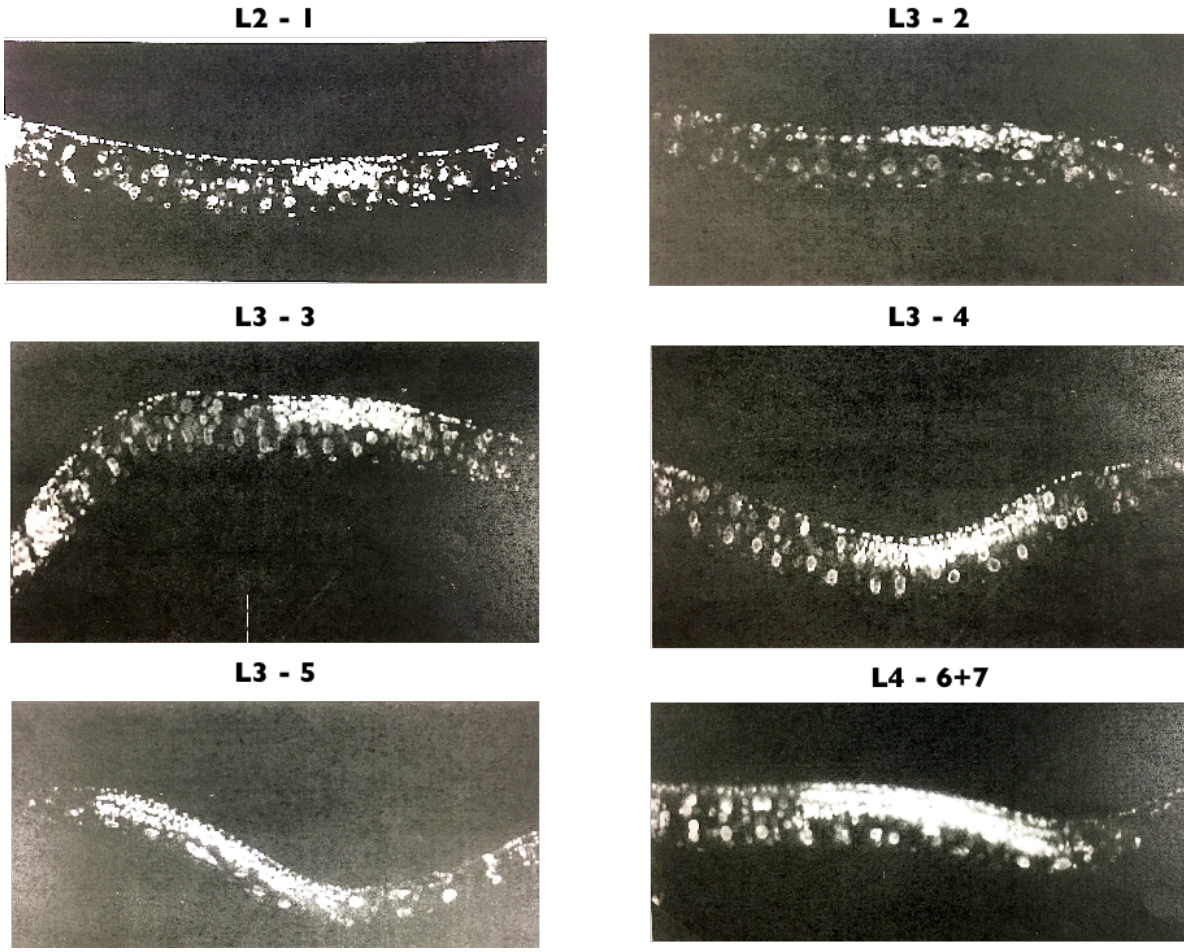
Worm culture and collection

On plate *E. coli* OP50 was poured onto Nematode Growth Medium (NGM) agar plates as a lawn and was allowed to grow for at least 3 days before use. Worms, especially L1 stage larvae, were either placed near the bacterial patch if transferred individually by a steel pick or added directly onto the patch if transferred as a suspension. Once gravid, worms were washed off the plate with M9 buffer (3 g KH_2PO_4, 6 g Na_2HPO_4, 5 g NaCl, 1 ml 1M $MgSO_4$, H_2O to 1L) and embryos were scrapped off with a L-shaped spreader if present. The solution was topped up with M9 buffer to 3 ml; 3 ml of 2x bleach solution (6 ml H_2O, 1 ml 10M NaOH, 3 ml 12% sodium hypochlorite) was added and the entire solution was inverted frequently for 5 to 7 min until most of the worms are dissolved. The embryos were washed thrice with 14 ml of M9 buffer by centrifuging at 2,000 rpm for 1 min. They were then left in 10 ml of M9 buffer and left overnight in a shaker to hatch at an appropriate temperature depending on the strain involved.

Liquid culture Worms were grown at a density of 8,000 worms/ml to avoid overcrowding in S Medium (1L S Basal [5.85 g NaCl, 1 g K_2HPO_4, 6 g KH_2PO_4, 1

ml cholesterol, H_2O to 1L], 10 ml 1M K citrate pH 6.0, 10 ml trace metals solution [1.86 g di-sodium EDTA, 0.69 g $FeS0_4$, 0.2 g $MnCl_2$, 0.29 g $ZnSO_4$, 0.025 g $CuSO_4$, H_2O to 1L], 3 ml 1M $CaCl_2$, 3 ml 1M MgSO4, 1x penicillin/streptomycin, 1x amphotericin B) at an appropriate temperature at 180 rpm in a shaker. When gravid, 5 ml of packed worms (sedimentation by gravity) was washed with M9 buffer and resuspended in 15 ml. They were then bleached with 15 ml of 2x bleach solution for 5 to 7 min. Depending on the amount of worm debris and the presence of dauer larvae (worms in stasis due to stress), embryos were cleaned with the sucrose float technique: 7 ml of embryos in M9 buffer with 7 ml of 60% sucrose solution, then centrifugation at 3,500 rpm for 3 min at 4C; embryos floating on the top were washed thrice with M9 buffer. To let the embryos hatch, they were left in M9 buffer or S Medium in a shaker at around 10 to 20 worms/ul.

Staging The L3 stage lasts for around 7 hours at 20C wherein the germline extends substantially. This extension can be separated into overlapping stages (1 to 6) and I chose 3 as the midpoint between L3 and L4 stages. To stain and visualise DNA, aliquots from worm collections were fixed in methanol for at least an hour. Samples were then resuspended in 1x PBS with 100 ng/ml of DAPI and incubated at RT for about 30 min. They were then washed thrice in 1x PBS and mounted for observation directly onto glass slides. All samples used in this thesis were staged and results are presented in **Appendix A2.2**.



Extension of germline during development.

Cell culture - S2 & GM12878

Drosophila S2 cells were grown in 15 ml of complete media (Schneider's *Drosophila* Medium, 10% heat-inactivated fetal bovine serum, 1x penicillin/ streptomycin) at 26C without CO_2 in 75 cm^2 flasks. Cells were seeded at a minimum of 2,000,000 cells/ml and passaged when cell density reached 10,000,000 cells/ml. Collections were made by centrifugation at 1,000*g* for 3 min.

GM12878 cells were grown in 15 ml of media (RPMI 1640, 2 mM L-glutamine, 50U/ml penicillin, 50 ug/ml streptomycin, 15% fetal bovine serum) at 37C under 5% CO_2 in 75 cm^2 flasks. They were seeded at a cell density of at least 200,000 cells/ml and passaged when they are at a density of 800,000 cells/ml. For experiments, cells were collected by spinning at 500*g* for 3 min at 4C when they are approximately at a density of 800,000 cells/ml.

Once a month, for both *Drosophila* S2 and GM12878, mycoplasma contamination was tested with the LookOut Mycoplasma qPCR Detection Kit. Fungal and bacterial contamination were tested visually whenever the opportunity arose.

Nuclei isolation of L3 stage *C. elegans* larvae

1 ml of L3 stage worm popcorn was crushed either by 2 strokes of the mechanical gun or 15 s in the grinder at 25 cycles/s. If fixation is required, formaldehyde was added at a ratio of 20 ml of formaldehyde to 1 ml of ground worm popcorn. Fixation was quenched by adding glycine to a final concentration of 125 mM and incubating for 5min at RT. Worm fragments were washed thrice with Buffer A by using a tabletop centrifuge at 2000 *rpm* for 1 min. After the washes, the worm pellet was resuspended in 7 ml of working Buffer A (+0.025% Igepal CA630) and incubated on ice for 15 min. The entire solution was

transferred to a steel dounce homogeniser and dounced for 20 times. To remove debris, the solution was spun at 100*g* for 6 min. The supernatant containing the nuclei was kept and the pellet was resuspended in 7 ml of working Buffer A. Douncing and debris removal were repeated and the supernatant was pooled with the nuclei from earlier. Nuclei were then stained with DAPI and counted using a C-Chip haemocytometer. For downstream experiments, the nuclei were aliquoted and spun down at 1,000*g* for 10 min.

ATAC-seq

1 million *C. elegans* nuclei were resuspended in 50 ul of tagmentation master-mix (25 ul 2x Nextera TD buffer, 22.5 ul H_2O, 2.5 ul Nextera Tn5) and incubated for 30 min at 37C with light shaking (400 rpm). The reaction was stopped with 5 ul of 1% SDS and tagmented DNA was purified with the use of a Qiagen MinElute column and eluted in 20 ul of Elution Buffer. 1 ul of the tagmented DNA was used to quantify the number of cycles needed for amplification via qPCR (typically 14-16 cycles) and the rest of the DNA was amplified, cleaned, and size-selected (200-500bp).

ARC-C

10 million fixed nuclei were resuspended in (200-y) ul of 1x DNase buffer (Roche), where y represents the amount of DNase I. Typically, 3 to 5 different

DNase I digestion concentrations are used to account for day-to-day variations in digestion levels. After nuclease digestion at 25C for 10 min, the reaction was quenched by a final concentration of 25 mM EDTA and 5 mM Tris pH 7.5. 5 ul of the nuclei was aliquoted, 0.25 ul of Proteinase K (NEB: 800 units/ml) was added and the suspension was incubated at 65C for at least 4 h for reverse cross-linking; the level of digestion was assayed by running the aliquot on a genomic ScreenTape. The rest of the nuclei was washed twice with 1 ml of ice-cold Nuclear Washing Buffer at a centrifugation speed of 1,000 *g* for 10 min at 4C. The nuclei were resuspended in 100 ul of end repair master-mix (10 ul 10x NEB End Repair Buffer, 5 ul NEB End Repair Enzyme Mix, 85 ul H_2O) and incubated at 20C for 30 min with shaking at 400 rpm in a table-top thermo-mixer. Thereafter, 400 ul of ligation master-mix was added (40 ul 10x ligation buffer, 5 ul T4 DNA ligase [400,000 units/ml], 355 ul H_2O) and the mixture was left on a rotating wheel at 16C for 4 h or 4C overnight. Nuclei were pelleted, resuspended in 50 ul of tagmentation master-mix (22.5 ul H_2O, 25 ul 2x Nextera TD buffer, 2.5 ul Nextera Tn5 transposase) and incubated at 37C for 30 min. 5 ml of 1% SDS and 2 ul of Proteinase K were added and left at 65C for 15 min before DNA is purified using Qiagen MinElute columns and eluted in 50 ul of Qiagen Elution Buffer. To obtain DNA fragments below 600 bp, I applied two rounds of size-selection with 0.6 volume AMPure XP beads. The resultant DNA was amplified and size-selected again with AMPure XP beads to a range of 200-700bp before being sequenced

paired-end. As quality control, in the case of *C. elegans* libraries, I ran quantitative PCR (qPCR) on a diagnostic DHS.

Primers for enriched DHS:

DHS01_F: GACGCATATTATTACACCCACGC
DHS01_R: GTGATTCGTGGTAGAGACGCA

Primers for background:
DHS05_F: ACATGGCTGGAAATTGGGGG
DHS05_R:GCGAACCCAATTTTGCGGAG

Poly(A) mRNA-seq

1 ml of Trizol was added to 50-100 ul of packed worms or 1 worm popcorn and the entire suspension was passed through a 27G needle at least 10 times. About 100 ul of autoclaved sand was added and the mixture was vortexed at max speed for 30 min with a tabletop vortexer. 200 ul of chloroform was added, vortexed for 15 s and incubated for 15 min at RT. To separate the layers containing DNA, proteins, or RNA, the solution was spun for 15 min at 12,000 rpm at 4C. The upper aqueous layer (~500 ul) containing RNA was transferred to a new micro-centrifuge tube; 1 ul of GlycoBlue co-precipitant and 50 ul of 3M sodium acetate to help precipitate RNA were added. Thereafter, 500 ul of isopropanol was added, the entire solution was inverted 6 times and left at RT for 10 min. It was then spun at 12,000 rpm at 4C for 10 min to pellet RNA, which was then washed in 1 ml of

75% ethanol. RNA was re-pelleted at 7,500 rpm at 4C for 5 min and the supernatant was removed. The pellet was air-dried for 5-10 min and resuspended in 87.5 ul DEPC water and incubated at 37C for 15 min. To remove DNA contamination, 10 ul of 10x DNase I buffer and 2.5 ul Baseline-ZERO DNase were added to the RNA. RNA cleanup was done with the RNeasy Mini Kit, eluted in 20 ul DEPC water and quantified using a Qubit fluorometer. Poly(A) mRNA-seq libraries were prepared with 100 ng to 1 ug of total RNA using a TruSeq RNA Library Prep Kit (v2).

CRISPR-Cas9 genome editing

CRISPR-Cas9 genome editing was used to generate the following strains: JA1799, JA1800, JA1801, JA1802. Injections were performed using the *dpy-10* co-CRISPR method to enrich for the desired edit (Arribere *et al* 2014, Paix *et al* 2015). The *dpy-10* repair template introduces a substitution that induces the dumpy phenotype; the presence of the *dpy-10* conversion implies that the locus of interest has been edited as well. Cas9 protein was made in-house according to Paix *et al* (2015). tracRNAs and crRNAs were purchased from Dharmacon; repair templates were purchased from IDT as Ultramer oligonucleotides (JA1798, JA1811, JA1850, JA1851). crRNAs were designed using the online CRISPOR tool (Haeussler *et al* 2016) to target two regions flanking the edit of interest. The DNA breaks were then repaired with the repair template by homologous recombination.

Each 10 ul of injection mix contained 10.5 uM Cas9 protein, 50 mM KCl, 100 mM HEPES pH 7.4, 56 uM tracRNA, 40 uM targeted gene crRNAs, 24 uM *dpy-10* crRNA, 0.88 uM *dpy-10* ssODN repair template, 7 uM ssODN targeted gene repair template. Constructs were injected into the syncytium of N2 Bristol *C. elegans* young adults. Injected worms were transferred onto individual NGM plates seeded with a lawn of OP50 and incubated at 25C. After 3 days, F1 progenies were scored for dumpy and roller phenotypes. All worms from two "jackpot broods" (Arribere *et al* 2014) (i.e. enriched number of dumpy and rollers; typically 40-60% dumpy and 5-10% rollers) were cloned onto individual plates and genotyped for the edit of interest. To revert the dumpy phenotype, which is recessive homozygous, back to wild-type whilst keeping the edit of interest, dumpy worms were crossed with N2 males; rollers are dominant recessive and can be removed by self-fertilisation.

crRNA	Sequence
Hub02_left	UCGACACGACAAUUUUUCUU
Hub02_right	CAUCUUUAGAAUGGUCAGUC
Hub03_left	CUCGGCUCAACGCGGCUUUA
Hub03_right	AUCAUGGGGUUUUCAAAGAA
Hub05_left	CACGUAUCGGUUCGAGCGCU
Hub05_right	GAAGUACAUCCCACUUUUGU

Repair template	Sequence
Hub02 ssODN	GTTACAGTACTCTTTAAAGGAGCATTTTTCCAAAGTGACCATTCT AAAGATGAAACCATGGTTACTGTAAG
Hub03 ssODN	CAAATATCCAGGTACTCCAGGGTACTCGGCTCAACGCGGCTGAA TGGATAGTATTTTGAAGCTTAAAATAAAAGTCG
Hub05 ssODN	GAAAAGAGAAGAATAGAACAGAAACTGAAACTAACCAAGTGT GGGAAAAAGAAAACTAATAAGTTCCTCGAAAC

Primer name	sequence
Hub02_out_FW	GTGCGCCTTTAAAGAGTATTG
Hub02_out_RV	GTTAGGTCTCAGCACGAAATC
Hub02_in_RV	CATACACTCACCATCAAATTGA
Hub03_out_FW	GATTTTACAGTATTCTAGGGTAG
Hub03_out_RV	CCTACAAAATGTGTGATGGTTT
Hub03_in_RV	CTAAATGACCTATTGTGGCAAG
Hub05_out_FW	CTTTCTATCTACATAGGTAGTTA
Hub05_out_RV	CATGCAATTTGTATAATACCACT
Hub05_in_RV	CGTTTGTTTATCGTCTTCTCC

Strains used

Strains	Genotype
N2	
YJ55	*blmp-1 (tm548) I.*
GW637	*met-2 (n4256) III; set-25 (n5021) III.*
JA1802 (hub02)	*chd-7 (we27) I.*

JA1808 (hub03)	*bath-43 (we26) III.*
JA1800 (hub05)	K04B12.2 (*we25*) *II.*
MT13954	*mir-81 & mir-82* (nDf54) *X.*
MT16494	*mir-229 & mir-64 & mir-65 & mir-66* (nDf63) *III.*
MT17429	nDf67 *IV.*
ST36	*plx-1 (nc36) IV.*

Processing ARC-C data

Adapter sequences were trimmed by **cutadapt** and sequences with under 20bp remaining were removed. Each sequenced ends were aligned independently to the ce10 reference genome by **BWA mem** which allows split-read alignment using the default parameters. The aligned ends (or their 5′ segment if the end produces split alignments) were then paired. We required both ends of a pair to align uniquely and with high confidence (mapping quality ≥ 30 and number of mismatches ≤ 2) to the nuclear genome and outside modENCODE backlisted-regions. PCR duplicates were next removed by **sambamba markdup**. The remaining read pairs were regarded as valid read pairs. Valid read pairs mapping to different chromosomes or greater than 600bp apart on the same chromosome were regarded as *trans*- or *cis*-informative read pairs, respectively. The 600bp threshold was established by comparing the proportions of the four possible end alignment orientation configurations (forward-forward, forward-reverse, reverse-forward and reverse-reverse) as a function of mapping distance. The forward-

reverse configuration (non-ligated fragments) was the vast majority under about 500bp, whereas above 600bp the proportions were stably the same at 25% each (**Fig 1.5**).

Contact maps were made from informative read pairs by binning the genome into fixed-width non-overlapping bins and counting the number of read pairs between each pair of bins. The maps were then normalised by matrix balancing using the Knight-Ruiz algorithm. For aggregated contact analysis, the map was further divided by the average contact frequency given distance from the diagonal to remove background slope of contact frequency.

Aggregated contact analysis

A contact is defined as a pair of genomic locations. Aggregated contact analysis is a method of visualising contact frequency of a group of many contacts with their local backgrounds to reveal average contact property of the group. We applied this method to both point genomic locations, such as transcription factor binding sites, and larger intervals, such as chromatin domains. For both analysis, we used matrix- balanced contact contact maps with distance-dependent background removed. In TFBS analysis, we used contact maps of 1kb resolution. For each TF, up to 10,000 contacts were randomly sampled from all possible *cis*-contacts among its binding sites within a distance range of 20kb to 1Mb, and local

maps of 21 by 21 bins centering at the contacts were extracted and aggregated. In domain analysis, we used contact maps of 5kb resolution. For each type of domain, we generated two set of contacts: a) intra-domain contacts, and b) all possible pairs of inter-domain contacts in the range of 50kb to 2Mb. Local maps of each contact with up to 25kb flanking interval (unless it hits the next domain border where it stops) each side were extracted and aggregated. The aggregated contact was scaled to a square of 10 by 10 bins and the flanking interval was scaled to 5 bins wide.

Calling significant interactions

We took a set of 42,245 annotated regulatory elements each sizing 150bp derived from multi-stage ATAC- seq and short-cap RNA-seq data. Elements within 100bp of each other were merged. The resulting intervals were expanded to 500bp or until neighbouring intervals began to touch. The rest of the genome were covered with evenly placed 500bp non-overlapping fixed-width intervals, hence the entire genome was covered by a combined set of 192,257 intervals (**I**).

For every interval $i \in \mathbf{I}$, the number of *cis*-informative read pairs $c_{cis,i}$ were counted. Intervals in the top 10% of the coverage distribution were regarded as peaks and intervals in the bottom 10% were removed. An off-peak *cis*-informative coverage $c_{offpeak,i}$ was calculated for every kept interval, counting the number of

contacts not involving peak intervals. We calculated a scaling factor for the interval's visibility as $v_i = (c_{offpeak,i}/\text{median}(c_{offpeak,.}))^{0.87}$. Chromosome-wide average distance-dependent contact frequency F(*d*) in the distance range of 1kb to 1Mb was modeled by fitting a spline function in a two-pass process. For every pair of intervals with a distance between 1kb to 1Mb, an expected contacts frequency was calculated given the distance and the visibility of each interval as $f_{i,j} = v_i v_j F(d_{i,j})$. Given the total number of *cis*-informative contacts (*N*) of the chromosome, we considered a null distribution in the form of a binomial, where the observed number of contacts, $n_{i,j} \sim \text{binomial}(N, f_{i,j})$. Significant interactions were called at an FDR level of 0.05 and were post-filtered requiring support by more than 5 read pairs.

REFERENCES

Adey, A. et al., 2010. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biology*, 11(12), p.R119.

Albiez, H. et al., 2006. Chromatin domains and the interchromatin compartment form structurally defined and functionally interacting nuclear networks. *Chromosome Research*, 14(7), pp.707–733.

Albritton, S.E. & Ercan, S., 2018. Caenorhabditis elegans Dosage Compensation: Insights into Condensin- Mediated Gene Regulation. *Trends in Genetics*, 34(1), pp. 41–53.

Albritton, S.E. et al., 2017. Cooperation between a hierarchical set of recruitment sites targets the X chromosome for dosage compensation. *eLife*, 6, p.9244.

Alipour, E. & Marko, J.F., 2012. Self-organization of domain structures by DNA-loop- extruding enzymes. *Nucleic Acids Research*, 40(22), pp.11202–11212.

Allan, J. et al., 2012. Micrococcal Nuclease Does Not Substantially Bias Nucleosome Mapping. *Journal of Molecular Biology*, 417(3), pp.152–164.

Altun, Z.F. & Hall, D.H., 2009. Introduction. *WormAtlas*. doi:10.3908/wormatlas. 1.1

Amidzadeh, Z. et al., 2014. Assessment of different permeabilization methods of minimizing damage to the adherent cells for detection of intracellular RNA by flow cytometry. *Avicenna journal of medical biotechnology*, 6(1), pp.38–46.

Amini, S. et al., 2014. Haplotype-resolved whole-genome sequencing by contiguity- preserving transposition and combinatorial indexing. *Nature Genetics*, 46(12), pp.1343– 1349.

Andersen, E.C. & Horvitz, H.R., 2007. Two C. elegans histone methyltransferases repress lin-3 EGF transcription to inhibit vulval development. *Development*, 134(16), pp.2991– 2999.

Andrey, G. & Mundlos, S., 2017. The three-dimensional genome: regulating gene expression during pluripotency and development. *Development*, 144(20), pp.3646–3658.

Andrey, G. et al., 2013. A switch between topological domains underlies HoxD genes collinearity in mouse limbs. *Science*, 340(6137), 1234167.

Andrey, G. et al., 2017. Characterization of hundreds of regulatory landscapes in developing limbs reveals two regimes of chromatin folding. *Genome Research*, 27(2), pp.223–233.

Arrigoni, L. et al., 2016. Standardizing chromatin research: a simple and universal method for ChIP-seq. *Nucleic Acids Research*, 44(7), pp.e67–e67.

Ay, F., Bailey, T.L. & Noble, W.S., 2014. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Research*, 24(6), pp.999–1011.

Bacher, C.P. et al., 2006. Transient colocalization of X-inactivation centres accompanies the initiation of X inactivation. *Nature Cell Biology*, 8(3), pp.293–299.

Bantignies, F. et al., 2011. Polycomb-Dependent Regulatory Contacts between Distant Hox Loci in Drosophila. *Cell*, 144(2), pp.214–226.

Barnes, T.M. et al., 1995. Meiotic recombination, noncoding DNA and genomic organization in Caenorhabditis elegans. *Genetics Society of America*, 141(1), pp.159–179.

Belyaeva, A. et al., 2017. Network analysis identifies chromosome intermingling regions as regulatory hotspots for transcription. *PNAS*, 114(52), pp.13714–13719.

Bessler, J.B., Andersen, E.C. & Villeneuve, A.M., 2010. Differential Localization and Independent Acquisition of the H3K9me2 and H3K9me3 Chromatin Modifications in the Caenorhabditis elegans Adult Germ Line. *PLoS Genetics*, 6(1), pp.e1000830.

Boettiger, A.N. et al., 2016. Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. *Nature*, 529(7586), pp.418–422.

Bohla, D. et al., 2014. A Functional Insulator Screen Identifies NURF and dREAM Components to Be Required for Enhancer-Blocking. *PLoS ONE*, 9(9), pp.e107765–13.

Bolzer, A. et al., 2005. Three-Dimensional Maps of All Chromosomes in Human Male Fibroblast Nuclei and Prometaphase Rosettes. *PLoS Biology*, 3(5), pp.e157–17.

Bothma, J.P. et al., 2015. Enhancer additivity and non-additivity are determined by enhancer strength in the Drosophila embryo. *eLife*, 4, p.1074.

Bouwman, B.A.M. & de Laat, W., 2015. Architectural hallmarks of the pluripotent genome. *FEBS Letters*, 589(PA), pp.2905–2913.

Boveri, T, 1909. Die Blastomerenkerne von *Ascaris megalocephala* und die Theorie der Chromosomeindividualität. *Arch Zellforsch*, 3, pp.181-268.

Boyle, A.P. et al., 2008. High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell*, 132(2), pp.311–322.

Boyle, S. et al., 2011. Fluorescence in situ hybridization with high-complexity repeat-free oligonucleotide probes generated by massively parallel synthesis. *Chromosome Research*, 19(7), pp.901–909.

Boyle, S. et al., 2001. The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Human molecular genetics*, 10(3), pp.211–219.

Brackley, C.A. et al., 2018. Extrusion without a motor: a new take on the loop extrusion model of genome organization. *Nucleus*, 9(1), pp.95-103.

Branco, M.R. & Pombo, A., 2006. Intermingling of Chromosome Territories in Interphase Suggests Role in Translocations and Transcription-Dependent Associations P. Becker, ed. *PLoS Biology*, 4(5), pp.e138–9.

Brejc, K. et al., 2017. Dynamic Control of X Chromosome Conformation and Repression by a Histone H4K20 Demethylase. *Cell*, 171(1), pp.85–102.e23.

Brown, J.M. et al., 2008. Association between active genes occurs at nuclear speckles and is modulated by chromatin environment. *The Journal of Cell Biology*, 182(6), pp.1083–1097.

Buenrostro, J.D. et al., 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 10(12), pp.1213–1218.

Cairns, J. et al., 2016. CHiCAGO: robust detection of DNA looping interactions in Capture Hi- C data. *Genome Biology*, pp.1–17.

Campbell, V.W. & Jackson, D.A., 1980. The effect of divalent cations on the mode of action of DNase I. The initial reaction products produced from covalently closed circular DNA. *Journal of Biological Chemistry*, 255(8), pp.3726–3735.

Cavalli, G., 2007. Chromosome kissing. *Current Opinion in Genetics and Development*, 17(5), pp.443–450.

Chambeyron, S. et al., 2005. Nuclear re-organisation of the Hoxb complex during mouse embryonic development. *Development*, 132(9), pp.2215–2223.

Chandler, M & Mahillion, J, 2002. Insertion sequences revisited. *Mobile DNA II*, pp305-366.

Chandra, T. et al., 2015. Global reorganization of the nuclear landscape in senescent cells. *CellReports*, 10(4), pp.471–483.

Chen, R.A.J. et al., 2014. Extreme HOT regions are CpG-dense promoters in C. elegans and humans. *Genome Research*, 24(7), pp.1138–1146.

Chen, X. et al., 2016. ATAC-see reveals the accessible genome by transposase-mediated imaging and sequencing. *Nature Methods*, 13(12), pp.1013–1020.

Cheutin, T. & Cavalli, G., 2014. Polycomb silencing: from linear chromatin domains to 3D chromosome folding. *Current Opinion in Genetics and Development*, 25, pp.30–37.

Clowney, E.J. et al., 2012. Nuclear Aggregation of Olfactory Receptor Genes Governs Their Monogenic Expression. *Cell*, 151(4), pp.724–737.

Collette, K.S. et al., 2011. Different roles for Aurora B in condensin targeting during mitosis and meiosis. *Journal of Cell Science*, 124(21), pp.3684–3694.

Corces, M.R. et al., 2017. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nature Methods*, 14(10), pp.959–962.

Corradin, O. et al., 2016. Modeling disease risk through analysis of physical interactions between genetic variants within chromatin regulatory circuitry. *Nature Genetics*, 48(11), pp.1313–1320.

Crane, E. et al., 2015. Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature*, 523(7559), pp.240–244.

Cremer, T. et al., 1982. Analysis of chromosome positions in the interphase nucleus of Chinese hamster cells by laser-UV-microirradiation experiments. *Human genetics*, 62(3), pp.201–209.

Cremer, T. et al., 2006. Chromosome territories – a functional nuclear landscape. *Current Opinion in Cell Biology*, 18(3), pp.307–316.

Criscione, S.W., De Cecco, M., et al., 2016. Reorganization of chromosome architecture in replicative cellular senescence. *Science Advances*, 2(2), pp.e1500882–13.

Criscione, S.W., Teo, Y.V. & Neretti, N., 2016. The Chromatin Landscape of Cellular Senescence. *Trends in Genetics*, 32(11), pp.751–761.

Croft, J.A. et al., 1999. Differences in the Localization and Morphology of Chromosomes in the Human Nucleus. *The Journal of Cell Biology*, 145(6), pp.1119–1131.

Csankovszki, G. et al., 2009. Three distinct condensin complexes control C. elegans chromosome dynamics. *Current biology : CB*, 19(1), pp.9–19.

Cui, M & Han, M, 2007. Roles of chromatin factors in *C. Elegans* development. *WormBook*. doi/10.1895/wormbook.1.139.1

Daugherty, A.C. et al., 2017. Chromatin accessibility dynamics reveal novel functional enhancers in C. elegans. *Genome Research*, 27(12), pp.2096–2107.

Davidson, I.F. et al., 2016. Rapid movement and transcriptional re-localization of human cohesin on DNA. *The EMBO Journal*, 35(24), pp.2671–2685.

Davies, J.O.J. et al., 2015. Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nature Methods*, 13(1), pp.74–80.

Davis, T.L. & Rebay, I., 2017. Master regulators in development: Views from the Drosophila retinal determination and mammalian pluripotency gene networks. *Developmental Biology*, 421(2), pp.93–107.

de Laat, W. & Duboule, D., 2013. Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature*, 502(7472), pp.499–506.

de Wit, E. et al., 2015. CTCF Binding Polarity Determines Chromatin Looping. *Molecular Cell*, 60(4), pp.676–684.

de Wit, E. et al., 2013. The pluripotent genome in three dimensions is shaped around pluripotency factors. *Nature*, 501(7466), pp.227–231.

Dekker, J. & Mirny, L., 2016. The 3D Genome as Moderator of Chromosomal Communication. *Cell*, 164(6), pp.1110–1121.

Dekker, J. et al., 2002. Capturing chromosome conformation. *Science*, 295(5558), pp.1306– 1311.

Dekker, J., Marti-Renom, M.A. & Mirny, L.A., 2013. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Reviews Genetics*, 14(6), pp.390–403.

Deng, W. et al., 2012. Controlling Long-Range Genomic Interactions at a Native Locus by Targeted Tethering of a Looping Factor. *Cell*, 149(6), pp.1233–1244.

Deng, X. et al., 2015. Bipartite structure of the inactive mouse X chromosome. *Genome Biology*, 16(1), pp.67–21.

Denholtz, M. et al., 2013. Long-Range Chromatin Contacts in Embryonic Stem Cells Reveal a Role for Pluripotency Factors and Polycomb Proteins in Genome Organization. *Cell Stem Cell*, 13(5), pp.602–616.

Denker, A. & de Laat, W., 2016. The second decade of 3C technologies: detailed insights into nuclear organization. *Genes & Development*, 30(12), pp.1357–1382.

Dixon, J.R. et al., 2015. Chromatin architecture reorganization during stem cell differentiation. *Nature*, 518(7539), pp.331–336.

Dixon, J.R. et al., 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398), pp.376–380.

Dorsett, D. & Merkenschlager, M., 2013. Cohesin at active genes: a unifying theme for cohesin and gene expression from model organisms to humans. *Current Opinion in Cell Biology*, 25(3), pp.327–333.

Dostie, J. et al., 2006. Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Research*, 16(10), pp.1299–1309.

Dowen, J.M. et al., 2014. Control of Cell Identity Genes Occurs in Insulated Neighborhoods in Mammalian Chromosomes. *Cell*, 159(2), pp.374–387.

Dryden, N.H. et al., 2014. Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome Research*, 24(11), pp.1854–1868.

Durand, N.C. et al., 2016. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell systems*, 3(1), pp.95–98.

Ebisuya, M. et al., 2008. Ripples from neighbouring transcription. *Nature Cell Biology*, 10(9), pp.1106–1113.

Eriksson, M. et al., 2003. Recurrent de novo point mutations in lamin A cause Hutchinson- Gilford progeria syndrome. *Nature*, 423(6937), pp.293–298.

Evans, K.J. et al., 2016. Stable Caenorhabditis elegans chromatin domains separate broadly expressed and developmentally regulated genes. *PNAS*, 113(45), pp. 7020-7029.

Falk, M. et al., 2018. Heterochromatin drives organization of conventional and inverted nuclei. *bioRxiv*, pp.1–19.

Fang, R. et al., 2016. Mapping of long-range chromatin interactions by proximity ligation- assisted ChIP-seq. *Nature Publishing Group*, 26(12), pp.1345–1348.

Farré, D. et al., 2007. Housekeeping genes tend to show reduced upstream sequence conservation. *Genome Biology*, 8(7), p.R140.

Fasulo, B. et al., 2012. The Drosophila Mi-2 Chromatin-Remodeling Factor Regulates Higher- Order Chromatin Structure and Cohesin Dynamics In Vivo. *PLoS Genetics*, 8(8), pp.e1002878–14.

Flavahan, W.A. et al., 2015. Insulator dysfuntion and oncogene activation in IDH mutant gliomas. *Nature*, 529(7584), pp.110–114.

Foley, J.W. & Sidow, A., 2013. Transcription-factor occupancy at HOT regions quantitatively predicts RNA polymerase recruitment in five human cell lines. *BMC Genomics*, 14(1), p.720.

Forcato, M. et al., 2017. Comparison of computational methods for Hi-C data analysis. *Nature Methods*, 14(7), pp.679–685.

Fortin, J.-P. & Hansen, K.D., 2015. Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biology*, 16(1), p. 180.

Franke, M. et al., 2016. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature*, 538(7624), pp.265–269.

Fraser, J., Ferrai, C., et al., 2015. Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Molecular Systems Biology*, 11(12), pp.852–852.

Fraser, J., Williamson, I., et al., 2015. An Overview of Genome Organization and How We Got There: from FISH to Hi-C. *Microbiology and Molecular Biology Reviews*, 79(3), pp.347– 372.

Fudenberg, G. et al., 2016. Formation of Chromosomal Domains by Loop Extrusion. *CellReports*, 15(9), pp.2038–2049.

Fullwood, M.J. et al., 2009. An oestrogen-receptor-α-bound human chromatin interactome. *Nature*, 462(7269), pp.58–64.

Ganji, M. et al., 2018. Real-time imaging of DNA loop extrusion by condensin. *Science*, 360(6384), pp.102–105.

Garrigues, J.M. et al., 2015. Defining heterochromatin in C. elegans through genome-wide analysis of the heterochromatin protein 1 homolog HPL-2. *Genome Research*, 25(1), pp.76– 88.

Gassler, J. et al., 2017. A mechanism of cohesin-dependent loop extrusion organizes zygotic genome architecture. *The EMBO Journal*, 36(24), pp.3600–3618.

Gavrilov, A.A., Golov, A.K. & Razin, S.V., 2013. Actual Ligation Frequencies in the Chromosome Conformation Capture Procedure A. Dean, ed. *PLoS ONE*, 8(3), p.e60403.

Gaydos, L.J. et al., 2012. Antagonism between MES-4 and Polycomb repressive complex 2 promotes appropriate gene expression in C. elegans germ cells. *CellReports*, 2(5), pp.1169– 1177.

Gerasimova, T.I., Byrd, K. & Corces, V.G., 2000. A chromatin insulator determines the nuclear localization of DNA. *Molecular Cell*, 6(5), pp.1025–1035.

Gerstein, M.B. et al., 2010. Integrative Analysis of the Caenorhabditis elegans Genome by the modENCODE Project. *Science*, 330(6012), pp.1775–1787.

Ghavi-Helm, Y. et al., 2014. Enhancer loops appear stable during development and are associated with paused polymerase. *Nature*, 513, p.89.

Gibcus, J.H. et al., 2018. A pathway for mitotic chromosome formation. *Science*, 359(6376), pp.eaao6135–19.

Giorgetti, L. & Heard, E., 2016. Closing the loop: 3C versus DNA FISH. *Genome Biology*, pp.1–9.

Goetze, S. et al., 2007. The Three-Dimensional Structure of Human Interphase Chromosomes Is Related to the Transcriptome Map. *Molecular and Cellular Biology*, 27(12), pp.4475–4487.

Goldman, R.D. et al., 2004. Accumulation of mutant lamin A causes progressive changes in nuclear architecture in Hutchinson-Gilford progeria syndrome. *Proceedings of the National Academy of Sciences*, 101(24), pp.8963–8968.

Goloborodko, A., Marko, J.F. & Mirny, L.A., 2016. Chromosome Compaction by Active Loop Extrusion. *Biophysj*, 110(10), pp.2162–2168.

Goryshin, I.Y. & Reznikoff, W.S., 1998. Tn5 in vitro transposition. *Journal of Biological Chemistry*, 273(13), pp.7367–7374.

Grasser, F. et al., 2008. Replication-timing-correlated spatial chromatin arrangements in cancer and in primate interphase nuclei. *Journal of Cell Science*, 121(11), pp.1876–1886.

Grob, S. et al., 2013. Characterization of chromosomal architecture in Arabidopsis by chromosome conformation capture. *Genome Biology*, 14(11), p.R129.

Grob, S., Schmid, M.W. & Grossniklaus, U., 2014. Hi-C Analysis in Arabidopsis Identifies the KNOT, a Structure with Similarities to the flamenco Locus of Drosophila. *Molecular Cell*, 55(5), pp.678–693.

Guo, Y. et al., 2015. CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell*, 162(4), pp.900–910.

Haarhuis, J.H.I. et al., 2017. The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. *Cell*, 169(4), pp.693–700.e14.

Hakimi, M.-A. et al., 2002. A chromatin remodelling complex that loads cohesin onto human chromosomes. *Nature*, 418(6901), pp.994–998.

Harewood, L. et al., 2017. Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours. pp.1–11.

Hay, D. et al., 2016. Genetic dissection of the α-globin super-enhancer in vivo. *Nature Genetics*, 48(8), pp.895–903.

Heger, P., Marin, B. & Schierenberg, E., 2009. Loss of the insulator protein CTCF during nematode evolution. *BMC Molecular Biology*, 10(1), pp.84–14.

Heitz, E, 1928. Das Heterochromatin der Moose. *Jahrb Wiss Botanik*, 69, pp.762-818.

Hendriks, G.-J. et al., 2014. Extensive Oscillatory Gene Expression during C. elegans Larval Development. *Molecular Cell*, 53(3), pp.380–392.

Hepperger, C. et al., 2008. Three-dimensional positioning of genes in mouse cell nuclei. *Chromosoma*, 117(6), pp.535–551.

Hnisz, D., Day, D.S. & Young, R.A., 2016. Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control. *Cell*, 167(5), pp.1188–1200.

Hnisz, D., Weintraub, A.S., et al., 2016. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science*, 351(6280), pp.1454–1458.

Ho, J.W.K. et al., 2014. Comparative analysis of metazoan chromatin organization. *Nature*, 512(7515), pp.449–452.

Horn, M. et al., 2014. DRE-1/FBXO11-Dependent Degradation of BLMP-1/ BLIMP-1 Governs C. elegans Developmental Timing and Maturation. *Developmental Cell*, 28(6), pp.697–710.

Hou, C. et al., 2012. Gene Density, Transcription, and Insulators Contribute to the Partition of the Drosophila Genome into Physical Domains. *Molecular Cell*, 48(3), pp.471–484.

Hsieh, T.-H.S. et al., 2015. Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell*, 162(1), pp.108–119.

Hu, M. et al., 2012. HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics*, 28(23), pp.3131–3133.

Huang, T.-F. et al., 2014. BLMP-1/Blimp-1 Regulates the Spatiotemporal Cell Migration Pattern in C. elegans A. D. Chisholm, ed. *PLoS Genetics*, 10(6), pp.e1004428–18.

Hug, C.B. et al., 2017. Chromatin Architecture Emerges during Zygotic Genome Activation Independent of Transcription. *Cell*, 169(2), pp.216–228.e19.

Hughes, J.R. et al., 2014. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nature Publishing Group*, 46(2), pp.205–212.

Hyun, M. et al., 2016. BLIMP-1/BLMP-1 and Metastasis-Associated Protein Regulate Stress Resistant Development in Caenorhabditis elegans. *Genetics Society of America*, 203(4), pp.1721–1732.

Ibn-Salem, J., Muro, E.M. & Andrade-Navarro, M.A., 2017. Co-regulation of paralog genes in the three-dimensional chromatin architecture. *Nucleic Acids Research*, 45(1), pp.81–91.

Ikegami, K. et al., 2010. Caenorhabditis elegans chromosome arms are anchored to the nuclear membrane via discontinuous association with LEM-2. *Genome Biology*, 11(12), p.R120.

Imakaev, M. et al., 2012. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature Methods*, 9(10), pp.999–1003.

Isoda, T. et al., 2017. Non-coding Transcription Instructs Chromatin Folding and Compartmentalization to Dictate Enhancer-Promoter Communication and T Cell Fate. *Cell*, 171(1), pp.103–119.e18.

Isono, K. et al., 2013. SAM Domain Polymerization Links Subnuclear Clustering of PRC1 to Gene Silencing. *Developmental Cell*, 26(6), pp.565–577.

Jänes, J. et al., 2018. Chromatin accessibility is dynamically regulated across C. elegans development and ageing. *bioRxiv*, pp.1–65.

Javierre, B.M. et al., 2016. Lineage-Specific Genome Architecture Links Enhancers and Non- coding Disease Variants to Target Gene Promoters. *Cell*, 167(5), pp. 1369–1384.e19.

Jäger, R. et al., 2015. Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nature Communications*, 6, p.6178.

Jin, F. et al., 2013. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, 489, p.57.

Jonkers, I. & Lis, J.T., 2015. Getting up to speed with transcription elongation by RNA polymerase II. *Nature Reviews Molecular Cell Biology*, 16(3), pp.167–177.

Junowicz, E. & Spencer, J.H., 1973. Studies on bovine pancreatic deoxyribonuclease A. I. General properties and activation with different bivalent metals. *Biochimica et biophysica acta*, 312(1), pp.72–84.

Kalhor, R. et al., 2011. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature Biotechnology*, 30(1), pp.90– 98.

Kamath, R.S. et al., 2003. Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature*, 421(6920), pp.231–237.

Ke, Y. et al., 2017. 3D Chromatin Structures of Mature Gametes and Structural Reprogramming during Mammalian Embryogenesis. *Cell*, 170(2), pp.367–381.e20.

Kharchenko, P.V. et al., 2010. Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature*, 471(7339), pp.480–485.

Kim, T.H. et al., 2007. Analysis of the Vertebrate Insulator Protein CTCF-Binding Sites in the Human Genome. *Cell*, 128(6), pp.1231–1245.

Kitagawa, R., 2009. Key players in chromosome segregation in Caenorhabditis elegans. *Frontiers in bioscience (Landmark edition)*, 14, pp.1529–1557.

Kolasinska-Zwierz, P. et al., 2009. Differential chromatin marking of introns and expressed exons by H3K36me3. *Nature Genetics*, 41(3), pp.376–381.

Kolovos, P. et al., 2014. Targeted Chromatin Capture (T2C): a novel high resolution high throughput method to detect genomic interactions and regulatory elements. *Epigenetics and Chromatin*, 7(1), p.10.

Kraft, K. et al., 2015. Deletions, Inversions, Duplications: Engineering of Structural Variants using CRISPR/Cas in Mice. *CellReports*, 10(5), pp.833–839.

Kranz, A.-L. et al., 2013. Genome-wide analysis of condensin binding in *Caenorhabditis elegans*. *Genome Biology*, 14(10), pp.R112–15.

Kudron, M.M. et al., 2018. The ModERN Resource: Genome-Wide Binding Profiles for Hundreds of Drosophila and Caenorhabditis elegans Transcription Factors. *Genetics Society of America*, 208(3), pp.937–949.

Kunitz, M., 1950. Crystalline desoxyribonuclease; isolation and general properties; spectrophotometric method for the measurement of desoxyribonuclease activity. *The Journal of general physiology*, 33(4), pp.349–362.

Küpper, K. et al., 2007. Radial chromatin positioning is shaped by local gene density, not by gene expression. *Chromosoma*, 116(3), pp.285–306.

Lajoie, B.R., Dekker, J. & Kaplan, N., 2015. The Hitchhiker's guide to Hi-C analysis: Practical guidelines. *Methods*, 72, pp.65–75.

Larkin, J.D., Papantonis, A. & Cook, P.R., 2013. Promoter type influences transcriptional topography by targeting genes to distinct nucleoplasmic sites. *Journal of Cell Science*, 126(9), pp.2052–2059.

Larson, A.G. et al., 2017. Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin. *Nature*, 22, pp.1–18.

LaSalle, J.M. & Lalande, M., 1996. Homologous association of oppositely imprinted chromosomal domains. *Science*, 272(5262), pp.725–728.

Latorre, I. et al., 2015. The DREAM complex promotes gene body H2A.Z for target repression. *Genes & Development*, 29(5), pp.495–500.

Le Gall, A., Valeri, A. & Nollmann, M., 2015. Roles of chromatin insulators in the formation of long-range contacts. *Nucleus*, 6(2), pp.118–122.

Lehner, B. et al., 2006. Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways. *Nature Genetics*, 38(8), pp.896–903.

Lercher, M.J., Blumenthal, T. & Hurst, L.D., 2003. Coexpression of neighboring genes in Caenorhabditis elegans is mostly due to operons and duplicate genes. *Genome Research*, 13(2), pp.238–243.

Lercher, M.J., Urrutia, A.O. & Hurst, L.D., 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nature Genetics*, 31(2), pp.180–183.

Lettice, L.A., 2003. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human molecular genetics*, 12(14), pp.1725–1735.

Li, R. et al., 1991. Direct interaction between Sp1 and the BPV enhancer E2 protein mediates synergistic activation of transcription. *Cell*, 65(3), pp.493–505.

Li, X. et al., 2017. Long-read ChIA-PET for base-pair-resolution mapping of haplotype- specific chromatin interactions. *Nature Protocols*, 12(5), pp.899–915.

Lieberman-Aiden, E. et al., 2009. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, 326(5950), pp.289–293.

Liu, T. et al., 2011. Broad chromosomal domains of histone modification patterns in C. elegans. *Genome Research*, 21(2), pp.227–236.

Lomvardas, S. et al., 2006. Interchromosomal Interactions and Olfactory Receptor Choice. *Cell*, 126(2), pp.403–413.

Lupiáñez, D.G. et al., 2015. Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell*, 161(5), pp.1012–1025.

Ma, H. et al., 1998. Spatial and temporal dynamics of DNA replication sites in mammalian cells. *The Journal of Cell Biology*, 143(6), pp.1415–1425.

Ma, W. et al., 2015. Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes. *Nature Methods*, 12(1), pp.71–78.

Ma, W. et al., 2018. Using DNase Hi-C techniques to map global and local three-dimensional genome architecture at high resolution. *Methods*, 142, pp.59–73.

Maniatis, T., Goodbourn, S. & Fischer, J.A., 1987. Regulation of inducible and tissue-specific gene expression. *Science*, 236(4806), pp.1237–1245.

Markenscoff-Papadimitriou, E. et al., 2014. Enhancer Interaction Networks as a Means for Singular Olfactory Receptor Expression. *Cell*, 159(3), pp.543–557.

Martin, P. et al., 2015. Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. *Nature Communications*, 6, pp.1–7.

McCord, R.P. et al., 2013. Correlated alterations in genome organization, histone methylation, and DNA-lamin A/C interactions in Hutchinson-Gilford progeria syndrome. *Genome Research*, 23(2), pp.260–269.

McGovern, A. et al., 2016. Capture Hi-C identifies a novel causal gene, IL20RA, in the pan- autoimmune genetic susceptibility region 6q23. *Genome Biology*, pp.1–15.

Meaburn, K.J. & Misteli, T., 2007. Cell biology: chromosome territories. *Nature*, 445(7126), pp.379–781.

Mendenhall, E.M. & Bernstein, B.E., 2008. Chromatin state maps: new technologies, new insights. *Current Opinion in Genetics and Development*, 18(2), pp. 109–115.

Mercer, T.R. & Mattick, J.S., 2013. Understanding the regulatory and transcriptional complexity of the genome through structure. *Genome Research*, 23(7), pp.1081–1088.

Meyer, B.J., 2010. Targeting X chromosomes for repression. *Current Opinion in Genetics and Development*, 20(2), pp.179–189.

Mifsud, B. et al., 2017. GOTHiC, a probabilistic model to resolve complex biases and to identify real interactions in Hi-C data. *PLoS ONE*, 12(4), pp.e0174744–15.

Mifsud, B. et al., 2015. Mapping long-range promoter contacts in human cells with high- resolution capture Hi-C. *Nature Genetics*, 47(6), pp.598–606.

Milani, P. et al., 2016. Cell freezing protocol suitable for ATAC-Seq on motor neurons derived from human induced pluripotent stem cells. *Scientific Reports*, pp. 1–10.

Mishra, A. & Hawkins, R.D., 2017. Three-dimensional genome architecture and emerging technologies: looping in disease. pp.1–14.

Misulovin, Z. et al., 2008. Association of cohesin and Nipped-B with transcriptionally active regions of the Drosophila melanogaster genome. *Chromosoma*, 117(1), pp.89–102.

Mitchell, J.A. & Fraser, P., 2008. Transcription factories are nuclear subcompartments that remain in the absence of transcription. *Genes & Development*, 22(1), pp.20–25.

Mito, Y., Sugimoto, A. & Yamamoto, M., 2003. Distinct Developmental Function of Two *Caenorhabditis elegans* Homologs of the Cohesin Subunit Scc1/Rad21. *Molecular Biology of the Cell*, pp.1–11.

Mizuguchi, T. et al., 2014. Cohesin-dependent globules and heterochromatin shape 3D genome architecture in S. pombe. *Nature*, 516(7531), pp.432–435.

Monahan, K. et al., 2017. Cooperative interactions enable singular olfactory receptor expression in mouse olfactory neurons. *eLife*, 6, p.1083.

Montefiori, L. et al., 2017. Reducing mitochondrial reads in ATAC-seq using CRISPR/Cas9. *Scientific Reports*, 7(1), pp.1213–9.

Morey, C., Kress, C. & Bickmore, W.A., 2009. Lack of bystander activation shows that localization exterior to chromosome territories is not sufficient to up-regulate gene expression. *Genome Research*, 19(7), pp.1184–1194.

Mumbach, M.R. et al., 2016. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nature Methods*, 13(11), pp.919–922.

Nagano, T. et al., 2017. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature*, 547(7661), pp.61–67.

Nagano, T. et al., 2015. Comparison of Hi-C results using in-solution versus in-nucleus ligation. *Genome Biology*, 16(1), p.1068.

Nakahashi, H. et al., 2013. A Genome-wide Map of CTCF Multivalency Redefines the CTCF Code. *CellReports*, 3(5), pp.1678–1689.

Nargund, A.M. et al., 2015. Mitochondrial and Nuclear Accumulation of the Transcription Factor ATFS-1 Promotes OXPHOS Recovery during the UPRmt. *Molecular Cell*, 58(1), pp.123–133.

Nargund, A.M. et al., 2012. Mitochondrial Import Efficiency of ATFS-1 Regulates Mitochondrial UPR Activation. *Science*, 337(6094), pp.587–590.

Narita, M. et al., 2003. Rb-Mediated Heterochromatin Formation and Silencing of E2F Target Genes during Cellular Senescence. *Cell*, 113(6), pp.703–716.

Nasmyth, K., Peters, J.M. & Uhlmann, F., 2000. Splitting the chromosome: cutting the ties that bind sister chromatids. *Science*, 288(5470), pp.1379–1385.

Naumann, T.A. & Reznikoff, W.S., 2000. Trans catalysis in Tn5 transposition. *Proceedings of the National Academy of Sciences*, 97(16), pp.8944–8949.

Naumova, N. et al., 2013. Organization of the Mitotic Chromosome. *Science*, 342(6161), pp.948–953.

Neusser, M. et al., 2007. Evolutionarily conserved, cell type and species-specific higher order chromatin arrangements in interphase nuclei of primates. *Chromosoma*, 116(3), pp.307–320.

Nichols, M.H. & Corces, V.G., 2015. A CTCF Code for 3D Genome Architecture. *Cell*, 162(4), pp.703–705.

Nolis, I.K. et al., 2009. Transcription factors mediate long-range enhancer-promoter interactions. *PNAS*, 106(48), pp.20222–20227.

Nora, E.P. et al., 2012. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, 485(7398), pp.381–385.

Nora, E.P. et al., 2017. Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell*, 169(5), pp. 930–933.e22.

Norton, H.K. & Phillips-Cremins J, 2017. Crossed wires: 3D genome misfolding in human disease. *Journal of Cell Biology*, 216(11), 3441.

Nuebler, J. et al., 2018. Chromatin organization by an interplay of loop extrusion and compartmental segregation. *Proceedings of the National Academy of Sciences*, 115(29), pp.E6697–E6706.

Ong, C.-T. & Corces, V.G., 2009. Insulators as mediators of intra- and inter-chromosomal interactions: a common evolutionary theme. *Journal of Biology*, 8(8), p.73.

Padeken, J. et al., 2013. The Nucleoplasmin Homolog NLP Mediates Centromere Clustering and Anchoring to the Nucleolus. *Molecular Cell*, 50(2), pp.236–249.

Padmanabha, D. et al., 2015. A HIF-independent mediator of transcriptional responses to oxygen deprivation in Caenorhabditis elegans. *Genetics Society of America*, 199(3), pp.739– 748.

Palstra, R.-J. et al., 2008. Maintenance of Long-Range DNA Interactions after Inhibition of Ongoing RNA Polymerase II Transcription. *PLoS ONE*, 3(2), pp.e1661–12.

Pant, V. et al., 2004. Mutation of a Single CTCF Target Site within the H19 Imprinting Control Region Leads to Loss of Igf2 Imprinting and Complex

Patterns of De Novo Methylation upon Maternal Inheritance. *Molecular and Cellular Biology*, 24(8), pp.3497– 3504.

Pennacchio, L.A. et al., 2013. Enhancers: five essential questions. *Nature Reviews Genetics*, 14(4), pp.288–295.

Peric-Hupkes, D. et al., 2010. Molecular Maps of the Reorganization of Genome-Nuclear Lamina Interactions during Differentiation. *Molecular Cell*, 38(4), pp.603–613.

Pernet, C.R. et al., 2013. Robust correlation analyses: false positive and power validation using a new open source Matlab toolbox. *Frontiers in Psychology*, 3, pp. 606.

Phanstiel, D.H. et al., 2017. Static and Dynamic DNA Loops form AP-1-Bound Activation Hubs during Macrophage Development. *Molecular Cell*, 67(6), pp. 1037-1048.

Phillips, J.E. & Corces, V.G., 2009. CTCF: Master Weaver of the Genome. *Cell*, 137(7), pp.1194–1211.

Phillips-Cremins, J.E. et al., 2013. Architectural Protein Subclasses Shape 3D Organization of Genomes during Lineage Commitment. *Cell*, 153(6), pp.1281–1295.

Picelli, S. et al., 2014. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Research*, 24(12), pp.2033–2040.

Pinkel, D. et al., 1988. Fluorescence in situ hybridization with human chromosome-specific libraries: detection of trisomy 21 and translocations of chromosome 4. *Proceedings of the National Academy of Sciences*, 85(23), pp.9138–9142.

Pirrotta, V. & Li, H.-B., 2012. A view of nuclear Polycomb bodies. *Current Opinion in Genetics and Development*, 22(2), pp.101–109.

Ponnaluri, V.K.C. et al., 2017. NicE-seq: high resolution open chromatin profiling. pp.1–15. Pope, B.D. et al., 2014. Topologically associating domains are stable units of replication-timing regulation. *Nature*, 515(7527), pp.402–405.

Prachumwat, A., DeVincentis, L. & Palopoli, M.F., 2004. Intron size correlates positively with recombination rate in Caenorhabditis elegans. *Genetics Society of America*, 166(3), pp.1585–1590.

Qiu, Z. et al., 2014. Functional Interactions between NURF and Ctcf Regulate Gene Expression. *Molecular and Cellular Biology*, 35(1), pp.224–237.

Quinodoz, S.A. et al., 2018. Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell*, 174(3), pp.744–757.e24.

Quintero-Cadena, P. & Sternberg, P.W., 2016. Enhancer Sharing Promotes Neighborhoods of Transcriptional Regulation Across Eukaryotes. *G3 (Bethesda, Md.)*, 6(12), pp.4167–4174.

Rabl, C, 1885. Über Zelltheilung. *Morph Jb*, 10, pp.214-330.

Racko, D. et al., 2017. Transcription-induced supercoiling as the driving force of chromatin loop extrusion during formation of TADs in interphase chromosomes. *Nucleic Acids Research*, 46(4), pp.1648–1660.

Ramani, V. et al., 2016. Mapping 3D genome architecture through in situ DNase Hi-C. *Nature Protocols*, 11(11), pp.2104–2121.

Ramírez, F. et al., 2015. High-Affinity Sites Form an Interaction Network to Facilitate Spreading of the MSL Complex across the X Chromosome in Drosophila. *Molecular Cell*, 60(1), pp.146–162.

Ramírez, F. et al., 2018. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nature Communications*, pp.1–15.

Rao, S.S.P. et al., 2014. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, 159(7), pp.1665–1680.

Rao, S.S.P. et al., 2017. Cohesin Loss Eliminates All Loop Domains. *Cell*, 171(2), pp. 305– 309.e24.

Ren, G. et al., 2017. CTCF-Mediated Enhancer-Promoter Interaction Is a Critical Regulator of Cell-to-Cell Variation of Gene Expression. *Molecular Cell*, 67(6), pp. 1049–1058.e6.

Reznikoff, W.S., 2008. Transposon Tn 5. *Annual Review of Genetics*, 42(1), pp.269–286.

Rizzino, A., 2009. Sox2 and Oct-3/4: a versatile pair of master regulators that orchestrate the self-renewal and pluripotency of embryonic stem cells. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 1(2), pp.228–236.

Robson, M.I. et al., 2017. Constrained release of lamina-associated enhancers and genes from the nuclear envelope during T-cell activation facilitates their association in chromosome compartments. *Genome Research*, 27(7), pp.1126–1138.

Rosa-Garrido, M. et al., 2017. High-Resolution Mapping of Chromatin Conformation in Cardiac Myocytes Reveals Structural Remodeling of the Epigenome in Heart Failure. *Circulation*, 136(17), pp.1613–1625.

Rudan, M.V. et al., 2015. Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *CellReports*, 10(8), pp.1297–1309.

Ryba, T. et al., 2010. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Research*, 20(6), pp.761–770.

Sahlén, P. et al., 2015. Genome-wide mapping of promoter-anchored interactions with close to single-enhancer resolution. *Genome Biology*, 16(1), pp.1–13.

Sanborn, A.L. et al., 2015. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences*, 112(47), pp.E6456–E6465.

Sanyal, A. et al., 2012. The long-range interaction landscape of gene promoters. *Nature*, 489(7414), pp.109–113.

Schaller, H, 1979. The intergenic region and the origins for filamentous phase DNA replication. *Cold Spring Harb Symp Quant Biol*, 43, pp.401-408.

Schmitt, A.D. et al., 2016. A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Reports*, 17(8), pp.2042–2059.

Schoenfelder, S. et al., 2018. Promoter Capture Hi-C: High-resolution, Genome-wide Profiling of Promoter Interactions. *Journal of Visualized Experiments*, (136), pp. 1–17.

Schoenfelder, S., Furlan-Magaril, M., et al., 2015a. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Research*, 25(4), pp.582–597.

Schoenfelder, S., Sugar, R., et al., 2015b. Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. *Nature Genetics*, 47(10), pp.1179– 1186.

Schwarzer, W. et al., 2017. Two independent modes of chromatin organization revealed by cohesin removal. *Nature*, 551(7678), pp.51–56.

Seddon, A.M., Curnow, P. & Booth, P.J., 2004. Membrane proteins, lipids and detergents: not just a soap opera. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1666(1-2), pp.105–117.

Sexton, T. et al., 2012. Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome. *Cell*, 148(3), pp.458–472.

Sofueva, S. et al., 2013. Cohesin-mediated interactions organize chromosomal domain architecture. *The EMBO Journal*, 32(24), pp.3119–3129.

Soler-Oliva, M.E. et al., 2017. Analysis of the relationship between coexpression domains and chromatin 3D organization X. *PLoS Computational Biology*, 13(9), pp.e1005708–25.

Song, L. & Crawford, G.E., 2010. DNase-seq: A High-Resolution Technique for Mapping Active Gene Regulatory Elements across the Genome from Mammalian Cells. *Cold Spring Harbor Protocols*, 2010(2), pp.pdb.prot5384–pdb.prot5384.

Spellman, P.T. & Rubin, G.M., 2002. Evidence for large domains of similarly expressed genes in the Drosophila genome. *Journal of Biology*, 1(1), p.5.

Stadler, M.R., Haines, J.E. & Eisen, M.B., 2017. Convergence of topological domain boundaries, insulators, and polytene interbands revealed by high-resolution mapping of chromatin contacts in the early *Drosophila* melanogaster embryo. *eLife*, 6, p.637662.

Stevens, T.J. et al., 2017. 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature*, 544(7648), pp.59–64.

Stigler, J. et al., 2016. Single-Molecule Imaging Reveals a Collapsed Conformational State for DNA-Bound Cohesin. *Cell Reports*, 15(5), pp.988–998.

Straub, T. et al., 2008. The Chromosomal High-Affinity Binding Sites for the Drosophila Dosage Compensation Complex. *PLoS Genetics*, 4(12), pp.e1000302–14.

Strom, A.R. et al., 2017. Phase separation drives heterochromatin domain formation. *Nature*, 547(7662), pp.241–245.

Szabo, Q. et al., 2018. TADs are 3D structural units of higher-order chromosome organization in Drosophila. *Science Advances*, 4(2), p.eaar8082.

Takebayashi, S.-I. et al., 2012. Chromatin-interaction compartment switch at developmentally regulated chromosomal domains reveals an unusual principle of chromatin folding. *PNAS*, 109(31), pp.12574–12579.

Tang, Z. et al., 2015. CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell*, 163(7), pp.1611–1627.

Terakawa, T. et al., 2017. The condensin complex is a mechanochemical motor that translocates along DNA. *Science*, 358(6363), pp.672–676.

Thakar, J. et al., 2015. Aging-dependent alterations in gene expression and a mitochondrial signature of responsiveness to human influenza vaccination. *Aging*, 7(1), pp.38–52.

Thierion, E. et al., 2017. Krox20 hindbrain regulation incorporates multiple modes of cooperation between cis-acting elements. *PLoS Genetics*, 13(7), pp.e1006903–18.

Thurman, R.E. et al., 2012. The accessible chromatin landscape of the human genome. *Nature*, 489(7414), pp.75–82.

Tolhuis, B. et al., 2002. Looping and interaction between hypersensitive sites in the active beta-globin locus. *Molecular Cell*, 10(6), pp.1453–1465.

Towbin, B.D. et al., 2012. Step-Wise Methylation of Histone H3K9 Positions Heterochromatin at the Nuclear Periphery. *Cell*, 150(5), pp.934–947.

Ulianov, S.V. et al., 2016. Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. *Genome Research*, 26(1), pp.70–84.

Van Bortle, K. et al., 2014. Insulator function and topological domain border strength scale with architectural protein occupancy. *Genome Biology*, 15(5), pp.R82–18.

van de Werken, H.J.G. et al., 2012. Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nature Methods*, 9(10), pp.969–972.

Victoria-Acosta, G. et al., 2015. Epigenetic silencing of the XAF1 gene is mediated by the loss of CTCF binding. *Scientific Reports*, pp.1–15.

Volpi, E.V. et al., 2000. Large-scale chromatin organization of the major histocompatibility complex and other regions of human chromosome 6 and its response to interferon in interphase nuclei. *Journal of Cell Science*, 113(9), pp.1565–1576.

Wang, Q. et al., 2017. Sub-kb Hi-C in D. melanogaster reveals conserved characteristics of TADs between insect and mammalian cells. *Nature Communications*, pp.1–8.

Wang, K. Et al., 1994. A simple method using T4 DNA polymerase to clone polymerase chain reaction products. *Biotechniques*, 17(2), pp236-238.

Wani, A.H. et al., 2016. Chromatin topology is coupled to Polycomb group protein subnuclear organization. *Nature Communications*, 7, pp.10291.

Webster, M., Witkin, K.L. & Cohen-Fix, O., 2009. Sizing up the nucleus: nuclear shape, size and nuclear-envelope assembly. *Journal of Cell Science*, 122(Pt 10), pp. 1477–1486.

Wei, Z. et al., 2013. Klf4 Organizes Long-Range Chromosomal Interactions with the Oct4 Locus in Reprogramming and Pluripotency. *Stem Cell*, 13(1), pp.36–47.

Weinreich, M.D., Gasch, A. & Reznikoff, W.S., 1994. Evidence that the cis preference of the Tn5 transposase is caused by nonproductive multimerization. *Genes & Development*, 8(19), pp.2363–2374.

Weintraub, A.S. et al., 2017. YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell*, 171(7), pp.1573–1588.e28.

Wiesenfahrt, T., Berg, J.Y., et al., 2016. The function and regulation of the GATA factor ELT-2 in the C. elegans endoderm. *Development*, 143(3), pp.483–491.

Wiesenfahrt, T., Osborne Nishimura, E., et al., 2016. Probing and rearranging the transcription factor network controlling the C. elegans endoderm. *Worm*, 5(3), p.e1198869.

Will, A.J. et al., 2017. Composition and dosage of a multipartite enhancer cluster control developmental expression of Ihh (Indian hedgehog). *Nature Genetics*, 49(10), pp.1539– 1545.

Williams, R.R.E. et al., 2002. Subchromosomal Positioning of the Epidermal Differentiation Complex (EDC) in Keratinocyte and Lymphoblast Interphase Nuclei. *Experimental Cell Research*, 272(2), pp.163–175.

Williamson, I. et al., 2014. Spatial genome organization: contrasting views from chromosome conformation capture and fluorescence in situ hybridization. *Genes & Development*, 28(24), pp.2778–2791.

Won, H. et al., 2016. Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature*, 538(7626), pp.523–527.

Wu, J. et al., 2016. The landscape of accessible chromatin in mammalian preimplantation embryos. *Nature*, 534(7609), pp.652–657.

Wutz, G. et al., 2017. Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *The EMBO Journal*, 36(24), pp.3573–3599.

Xiao, T., Wallace, J. & Felsenfeld, G., 2011. Specific Sites in the C Terminus of CTCF Interact with the SA2 Subunit of the Cohesin Complex and Are Required

for Cohesin-Dependent Insulation Activity. *Molecular and Cellular Biology*, 31(11), pp.2174–2183.

Xie, S. et al., 2017. Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. *Molecular Cell*, 66(2), pp.285–299.e5.

Xu, N., Tsai, C.-L. & Lee, J.T., 2006. Transient homologous chromosome pairing marks the onset of X inactivation. *Science*, 311(5764), pp.1149–1152.

Yaffe, E. & Tanay, A., 2011. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature Genetics*, 43(11), pp.1059–1065.

Yardimci, G.G. et al., 2018. Measuring the reproducibility and quality of Hi-C data. *BioRxiv*.

Yip, K.Y. et al., 2012. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biology*, 13(9), p.R48.

Yuen, K.C. & Gerton, J.L., 2018. Taking cohesin and condensin in context. *PLoS Genetics*, 14(1), pp.e1007118-14.

Yuen, K.C., Slaughter, B.D. & Gerton, J.L., 2017. Condensin II is anchored by TFIIIC and H3K4me3 in the mammalian genome and supports the expression of active dense gene clusters. *Science Advances*, 3(6), p.e1700191.

Zacher, B. et al., 2017. Accurate Promoter and Enhancer Identification in 127 ENCODE and Roadmap Epigenomics Cell Types and Tissues by GenoSTAN. *PLoS ONE*, 12(1), p.e0169249.

Zhan, Y. et al., 2017. Reciprocal insulation analysis of Hi-C data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes. *Genome Research*, 27(3), pp.479–490.

Zhang, Y. et al., 2012. Spatial Organization of the Mouse Genome and Its Role in Recurrent Chromosomal Translocations. *Cell*, 148(5), pp.908–921.

Zhao, Z. et al., 2006. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature Genetics*, 38(11), pp.1341–1347.

Zhou, M. & Reznikoff, W.S., 1997. Tn5 transposase mutants that alter DNA binding specificity. *Journal of Molecular Biology*, 271(3), pp.362–373.

Zhou, M., Bhasin, A. & Reznikoff, W.S., 1998. Molecular genetic analysis of transposase-end DNA sequence recognition: cooperativity of three adjacent base-pairs in specific interaction with a mutant Tn5 transposase. *Journal of Molecular Biology*, 276(5), pp.913–925.

Zhu, W. et al., 2017. Altered chromatin compaction and histone methylation drive non- additive gene expression in an interspecific Arabidopsis hybrid. pp.1–16.

Zuin, J. et al., 2014. Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proceedings of the National Academy of Sciences*, 111(3), pp.996–1001.

INDEX OF FIGURES

Figure 1.1	Aggregate distribution of restriction enzyme cut sites over transcription start sites in *C. elegans*	35
Figure 1.2	Count of DpnII restriction fragments with varying number of DHS in *Drosophila* S2, human K562, and *C. elegans* N2	37
Figure 1.3	Schematic of ARC-C protocol	39
Figure 1.4	Schematic of ARC-C data processing steps	41
Figure 1.5	Counts from library "N2_1" of inserts of each mapping orientation at varying insert sizes	42
Table 1.6	Key statistics for wild-type L3 libraries	44
Figure 1.7	Pearson correlation between wild-type L3 ARC-C library replicates binned at 1kb, 2kb, 5kb, 10kb, 20kb, and 50kb resolution	44
Figure 1.8	Coverage of HiCUP-filtered informative Hi-C reads from wild- type mixed Embryos, wild-type L3 stage ARC-C valid reads, wild-type L3 stage ATAC-seq normalised reads and genes	45
Figure 1.9	Aggregate coverage of N2_1a ARC-C valid reads and L3 ATAC-seq normalised reads over L3 ATAC-seq peaks	46
Figure 1.10	Gel electrophoresis of DNase I digested chromatin on 1% agarose gel	48
Table 1.11	Aggregate coverage enrichment of informative reads over DHS genome-wide of ARC-C libraries at different extents of DNase I digestion	48

Figure 1.12	Wild-type ATAC-seq coverage at different developmental stages	49
Figure 1.13	Genome-wide intra- and inter-chromosomal contact map for wild-type L3 stage C. elegans ARC-C	53
Figure 1.14	Log-log plot of informative ARC-C against informative Crane Hi-C reads binned at 50kb resolution	54
Figure 1.15	Comparison of ARC-C and Crane Hi-C contact map for chr I	55
Figure 1.16	Comparison of ARC-C and Crane Hi-C contact map for chr X	56
Figure 1.17	Comparison of ARC-C and Crane Hi-C insulation profiles across chr X	56
Figure 1.18	Circos plot of HiCUP-processed informative interactions from Crane Hi-C	58
Figure 1.19	Circos plot of informative interactions from ARC-C	59
Figure 1.20	Section from Circos plot in chr I: 3,706,000-3,745,000	60
Figure 1.21	Comparison of VC, Sqrtc, and KR on GM12878 Hi-C data	65
Figure 1.22	Quantifying APA scores or peak foci enrichment	66
Figure 1.23	Summary of ARC-C normalisation experiments	67
Figure 1.24	Breakdown of significant interactions by interaction-types and distances	69-71
Figure 1.25	Plot of correction coefficients derived from matrix balancing against various adjusted coverages	74
Figure 1.26	Informative coverage in DNase Hi-C and ARC-C centred around their corresponding hypersensitive sites	76
Figure 1.27	Summary of statistics in targeted DNase Hi-C and ARC-C	77

Figure 2.1A	Proportions of EE, PE, and PP interactions at different distance intervals	82
Figure 2.1B	Fold-change of observed over expected proportions for each interaction-type at different intervals	83
Figure 2.2	Expression correlation between promoter-promoter pairs at different distance intervals	85
Figure 2.3	Expression correlation of promoter-promoter pairs for All, HOT-HOT, HOT- nonHOT, and nonHOT-nonHOT classes	88
Figure 2.4	Gene expression CV or expression as a function of the number of enhancers	90
Figure 2.5	Fraction of genes with at least one promoter that have been grouped into deciles based on CV values	91
Figure 2.6	Aggregate coverage of L3 ATAC-seq centred on HOT, hubs, or L3 ATAC-seq peaks	93
Table 2.7	Summary of hub deletion RNA-seq experiments	94
Figure 2.8	IGV screenshots of CRISPR-Cas9 hub deletions - hub02, hub03, hub05	95
Figure 2.9	Principal component analysis of wild-type and hub deletion RNA-seq	96
Figure 2.10	Linked-genes analysis for hub02 deletion	99
Figure 2.11	Local genes analysis for hub02 deletions	100
Figure 2.12	Standard deviation of log2FC of *cis* informative peaks between hub05 and wild-type (N2), hub05 and hub05, wild-type and wild-type	103
Figure 2.13	Standard deviation of log2FC of *cis* informative peaks between hub02 and wild-type; hub03 and wild type	104

Figure 2.14	Wild-type and hub02 Circos plots	107-108
Figure 2.15	Zoomed in wild-type and hub02 Circos plots	109
Figure 2.16	Wild-type and hub03 Circos plots	110-111
Figure 2.17	Zoomed in wild -type and hub03 Circos plots	112
Figure 2.18	Wild-type and hub05 Circos plots	113-114
Figure 2.19	Zoomed in wild -type and hub05 Circos plots	115
Figure 3.1	Summary of factor APA in Crane Hi-C and ARC-C	120-121
Figure 3.2	Description and summary of factors that passed criteria	122-124
Figure 3.3	APA plots of KLE-2, PQN-85, and ELT-2 paired peaks within 20kb-1.5Mb in ARC-C and Crane Hi-C	124
Figure 3.4	Snapshot of contact map in mouse (chr8: 58,151,693-122,470,100)	127
Figure 3.5	Expression profiles of *scc-1*, *coh-1* and *kle-2* in different tissues	129-131
Figure 3.6	Correlation of KLE-2, COH-1, SCC-1 ChIP-seq at regulatory elements with at least 10 interactions	132
Figure 3.7	Heatmap of COH-1, SCC-1, KLE-2 in 2kb windows centred over L3 ATAC peaks.	132
Figure 3.8	Snapshot of significant interactions, COH-1 ChIP, SCC-1 ChIP, KLE-2 ChIP, chromatin state domains- active (red), regulated (black), genes	133
Figure 3.9	Snapshot of significant interactions, COH-1 ChIP, SCC-1 ChIP, KLE-2 ChIP, chromatin state domains- active (red), regulated (black), genes	134

Figure 3.10	Aggregate plot of z-scored COH-1 ChIP signal over pseudo-scaled active (green) or regulated (blue) chromatin state domains	134
Figure 3.11	Jaccard index heat map of condensin II subunits	137
Figure 3.12	Jaccard index heat map of chromatin regulators identified by factor APA	137
Figure 3.13	Factor APA of BLMP-1 and control factors ELT-2, KLE-2, SCC-1 in N2 and blmp-1 mutants at 1kb resolution	143
Figure 3.14	Log2-log2 plot of factor APA scores in wild-type (N2) over *blmp-1* mutants	144
Figure 3.15	Log2-log2 plot of factor APA scores in wild-type (N2) over *blmp-1* mutants	145
Figure 3.16	Log2-log2 plot of factor APA scores in wild-type (N2) over *blmp-1* mutants	146
Figure 4.1	Representative view of H3K36me3 and H3K27me3 forming broad expanses	148
Figure 4.2	Illustration of domain APA and compartment APA	149
Figure 4.3	Domain and compartment APA for active and regulated chromatin state domains in wild-type ARC-C	150
Table 4.4	Summary of compartment APA scores in Drosophila and C. elegans in active and inactive/regulated domains	153
Figure 4.5	Correlation of histone marks across human, fly, and worm	154
Figure 4.6	Distribution of H3K9me2/3 over chr I	155
Figure 4.7	Zoomed-in view of Figure 4.6	156

Figure 4.8	Domain and compartment APA for active, regulated, H3K9me2, H3K9me3 domains in wild-type ARC-C	157
Figure 4.9	Log2FC in accessibility between met-2 set-25 and wild-type worms	160
Figure 4.10	Log2FC in accessibility between met-2 set-25 and wild-type worms within active or regulated chromatin state domains	161
Figure 4.11	Log2FC in accessibility between met-2 set-25 and wild-type worms at different levels of wild-type H3K9me2 binding	162
Figure 4.12	Contact probability as a function of genomic distance; short	167
Figure 4.13	Contact probability as a function of genomic distance; long	168
Figure 4.14	Frequency of interactions in met-2 set-25 (metset) and wild-type (N2) at long, median, and short ranges separated by location	169
Figure 4.15	Domain and compartment APA for active and regulated chromatin state domains in wild-type (N2) and met-2 set-25	170
Figure 4.16	Coverage of H3K27me3 ChIP-seq over regulated, active chromatin state domains and borders in met-2 set-25 and wild-type (N2)	171
Figure 5.1	Gel electrophoresis of ~ 1ug DFF-digested chromatin on 1.5% agarose gel	174
Figure 5.2	Graphical schematic of ngARC-C protocol	176

Figure 5.3	Single-molecule imaging of YOYO-1 fluorescent dye labelled DNA after Tn5 transposition	178
Table 5.4	Key statistics for optimisation experiments, excluding ngARC-C	180
Table 5.5	Summary of ngARC-C optimisation experiments with key diagnostic statistics	183
Figure 5.6	Schematic of ngARC-C protocol and issues to resolve at corresponding steps	186
Figure 5.7	Gel electrophoresis of 100ng of DNA after Tn5 transposition on 1.5% agarose gel	187
Figure 5.8	Ratio of biotinylated to non-biotinylated oligonucleotides after pull-down, with or without pre-blocking	188
Figure 5.9	Relative abundance of mtDNA to nuclear DNA in untreated and CARM- depleted conditions	189
Figure A1.1	Summary of ARC-C library statistics	247
Figure A1.2	Staging for worm collections	248
Table A1.3	P-values for linked-genes analysis	249
Table A1.4	P-values for local-genes analysis	250
Figure A1.5	Linked-genes analysis for hub deletions	251-257
Figure A1.6	Local-genes analysis for hub deletions	258-265

LIST OF ABBREVIATIONS

APA: Aggregate Peak Analysis

ARC- C: Accessible Region Conformation-Capture

ATAC-seq: Assay for Transposase-Accessible Chromatin

BE: Boundary Element

BX-C: Bithorax-Complex

bZip: Basic Leucine Zipper

C-based methods: Chromosome conformation capture techniques

CAD: Caspase-Activated DNase

CAGE: Cap Analysis Gene Expression

CARM: Cas9-Assisted Mitochondrial DNA depletion

CHi-C: Capture-Hi-C

ChIA-PET: Chromatin Interaction Analysis with Paired-End-Tag sequencing

CHiCAGO: Capture Hi-C Analysis of Genomic Organisation

ChIP: Chromatin Immunoprecipitation

CNV-seq: Copy Number Variation using shotgun sequencing

CTCF: CCCTC-binding Factor

CT: Chromosome Territories

CV: Coefficient of Variance

DCC: Dosage Compensation Complex

DFF: DNA Fragmentation Factor

DHS: DNase I Hypersensitivity Site

DNase-seq: DNase I hypersensitive sites sequencing

DREAM: Dimerisation Partner, RB-like, E2F and Multi-vulval class B

EDTA: Ethylenediaminetetraacetic Acid

EE: Enhancer-Enhancer interactions

ENCODE: Encyclopaedia of DNA Elements

EPI: Enhancer-Promoter Interactions

ES: End Sequence

FC: Fold-Change

FDR: False Discovery Rate

FIREs: Frequently Interacting Enhancer Regions

FISH: Fluorescence in Situ Hybridization

GC-content: Guanine-Cytosine content

gDNA: genomic DNA

GFP: Green Fluorescent Protein

GWAS: Genome-Wide Association Studies

HAS: High Affinity Sites

HiCUP: Hi-C User Pipeline

HMM: Hidden Markov Modelling

HOT: High Occupancy Target

HP1: Heterochromatin Protein 1

H3K9me: H3K9 methylation

Ihh: Indian hedgehog gene

IncRNA: Long non-coding RNA

IS50: Insertion Sequence 50

LADs: Lamina-Associated Domains

LC-MRM: Liquid Chromatography-Multiple Reaction Monitoring mass spectrometry

LEF: Loop Extruding Factor

LOWESS: Locally Weighted Scatterplot Smoothing

MACS: Model-based Analysis for ChIP-Seq

MB: Matrix Balancing

ME: Mosaic ES

mESCs: mouse Embryonic Stem Cells

modENCODE: model organism ENCODE

mtDNA: Mitochondrial DNA

NADs: Nucleolar-Associated Domains

ncRNA: non-coding RNA

NEXSON: Nuclei Extraction by Sonication

ngARC-C: next-generation ARC-C

NuRD: Nucleosome Remodelling and Deacetylase

NURF: Nucleosome Remodelling Factor

OR: Olfactory Receptor

PCA: Principle Component Analysis

PcG: Polycomb Group

PCR: Polymerase Chain Reaction

PE: Promoter-Enhancer interactions

PLAC-seq: Proximity Ligation-Assisted ChIP-seq

PO: Promoter-Others

PP: Promoter-Promoter interactions

rex: recruitment element on the X

roX: RNA on the X chromosome

QB: Qiagen Buffer

qPCR: quantitative PCR

RNA pol II: RNA polymerase II

RNAi: RNA interference

SAHF: Senescence Associated Heterochromatic Foci

sci-RNA-seq: single-cell combinatorial indexing RNA sequencing

SDS: Sodium Dodecyl Sulphate

sgRNA: single guide RNA

SMC: Structural Maintenance of Chromosomes

Sqrtc: Square root of VC

TAD: Topologically Associating Domain

TF: Transcription Factor

TFIIIC: Transcription Factor IIIC

Tris-HCl: Tris Hydrochloride

TSS: Transcriptional Start Site

T2C: Targeted Chromatin Capture

UPRmt: mitochondrial Unfolded Protein Response

UPRmtE: UPRmt Element

VC: Vanilla Correction

3C: Chromosome Conformation Capture

4C: Circular Chromosome Conformation Capture

5C: Chromosome Conformation Capture Carbon Copy

APPENDIX A1 - SUPPLEMENTARY DATA

NameInThesis	Sample Type	DatasetFullName	nCleaned	nMapped	nRmBlacklistMT	nMq30Paire	nValid	nInfo	nCis	AtacPeakEnrichmer	CisRatio%	Complexity%
N2_1a	N2 L3	WQS014_wqs01.e4_N2_L3_DN-NT.v2_100u_run07c	359,454,110	339,865,256	322,494,860	245,647,846	109,921,232	9,321,648	7,253,030	3.7	77.81%	44.75%
N2_2a		WQS044_wqs05.e1_N2_L3_DN-NT.v2_50u_run11c	237,578,626	226,265,095	214,368,396	159,352,418	89,174,244	7,730,398	6,034,950	3.9	78.07%	55.96%
N2_2b		WQS062_wqs05.e2_N2_L3_DN-NT.v2_50u_run15c	247,833,286	241,537,115	227,665,927	180,531,150	134,038,220	8,137,112	6,336,580	3.4	77.87%	74.25%
N2_3		WQS117_wqs33.e1_N2_L3_DN-NT.v8_100u_run25c	25,401,652	24,758,121	23,364,577	18,570,720	17,692,004	987,838	769,404	4.91	77.89%	95.27%
N2_3		WQS122_wqs33.e1_N2_L3_DN-NT.v8_100u_run27c	64,660,032	63,046,826	55,628,990	43,229,076	19,532,052	1,579,066	1,239,860	5.38	78.52%	45.18%
N2_3		WQS123_wqs33.e1_N2_L3_DN-NT.v8_100u_run27c	88,372,016	86,138,430	76,020,428	59,119,348	24,897,760	1,994,554	1,565,798	5.39	78.50%	42.11%
N2_3		WQS124_wqs33.e1_N2_L3_DN-NT.v8_100u_run27c	45,466,636	44,333,253	39,055,393	30,365,534	18,454,360	1,507,336	1,185,686	5.5	78.66%	60.77%
bhmp-1	*bhmp-1* L3	WQS103_wqs14.e1_bhmp-1_L3_DN-NT.v6_50u_run20c	57,766,610	56,336,629	46,626,804	36,654,394	22,927,394	2,633,794	1,813,074	2.51	68.84%	62.55%
		WQS104_wqs14.e1_bhmp-1_L3_DN-NT.v6_100u_run20c	35,508,508	34,102,430	29,701,303	22,502,704	14,144,726	1,731,558	1,098,012	2.75	63.41%	62.86%
		WQS109_wqs15.e1_bhmp-1_L3_DN-NT.v6_50u_run21c	54,359,964	52,657,085	48,043,633	37,997,414	33,580,910	3,775,694	2,527,106	1.94	66.93%	88.38%
		WQS110_wqs15.e1_bhmp-1_L3_DN-NT.v6_100u_run21c	25,650,826	24,939,227	23,492,815	18,495,804	16,670,356	1,771,372	1,160,574	2.28	65.52%	90.13%
		WQS110_wqs15.e1_bhmp-1_L3_DN-NT.v6_100u_run22c	12,245,852	11,898,151	11,217,238	8,342,772	8,058,746	869,264	571,946	2.32	65.80%	96.60%
met-2 set-25	*met-2 set-25* L3	WQS100_wqs16.e1_metset_L3_DN-NT.v6_50u_run20c	88,468,280	85,819,074	71,836,217	52,390,216	41,599,528	3,517,620	2,876,078	3.33	81.76%	79.40%
		WQS101_wqs16.e1_metset_L3_DN-NT.v6_100u_run20c	37,471,238	36,063,750	32,128,931	24,105,622	15,864,820	1,329,962	985,746	4.64	74.12%	65.81%
		WQS107_wqs20.e1_metset_L3_DN-NT.v6_50u_run21c	55,018,564	54,000,265	38,695,459	30,539,936	24,569,344	1,964,000	1,443,196	3.84	73.48%	80.45%
		WQS108_wqs20.e1_metset_L3_DN-NT.v6_100u_run21c	32,140,272	31,316,905	23,625,525	16,358,782	14,770,438	1,381,354	990,152	3.6	71.68%	90.29%
		WQS107_wqs20.e1_metset_L3_DN-NT.v6_50u_run22c	24,319,838	23,863,866	17,179,480	12,639,806	11,696,022	982,796	742,082	3.66	75.51%	92.53%
		WQS107_wqs20.e1_metset_L3_DN-NT.v6_50u_run23c	129,376,498	126,966,219	91,894,595	67,793,656	47,164,700	3,553,934	2,642,480	3.84	74.35%	69.57%
		WQS101_wqs16.e1_metset_L3_DN-NT.v6_100u_run23c	44,591,544	42,899,498	38,116,517	26,899,740	18,464,904	1,387,418	1,031,772	4.78	74.37%	68.64%
hub02	hub02 L3	WQS120_wqs34.e1_hub02_L3_DN-NT.v8_100u_run27c	88,348,240	86,665,433	49,054,283	39,665,728	31,375,156	1,124,482	820,506	3.64	72.97%	79.10%
		WQS126_wqs36.e1_hub02_L3_DN-NT.v8_50u_run28c	111,019,398	108,828,514	60,669,467	48,974,772	35,925,538	1,285,528	937,600	3.67	72.94%	73.36%
		WQS131_wqs34.e1_hub02_L3_DN-NT.v8_50u_run31c	48,644,096	48,096,952	39,052,044	31,939,394	22,358,812	1,186,018	906,026	6.24	76.39%	70.00%
		WQS120_wqs34.e1_hub02_L3_DN-NT.v8_100u_run31c	174,533,896	171,299,903	95,963,177	77,561,056	49,119,442	1,779,154	1,297,506	3.73	72.93%	63.33%
		WQS126_wqs36.e1_hub02_L3_DN-NT.v8_50u_run31c	158,583,608	156,390,831	95,995,053	76,448,786	36,873,538	2,431,026	1,793,924	4.27	73.79%	48.23%
hub03	hub03 L3	WQS127_wqs37.e1_hub03_L3_DN-NT.v8_50u_run28c	92,267,188	90,643,149	75,206,782	62,357,552	41,254,930	2,659,776	2,188,674	6.98	82.29%	66.16%
		WQS128_wqs37.e2_hub03_L3_DN-NT.v8_50u_run31c	187,291,902	183,598,930	133,063,410	104,620,230	51,491,028	4,153,566	3,319,990	5.18	79.93%	49.22%
hub05	hub05 L3	WQS114_wqs23.e1_hub05_L3_DN-NT.v8_50u_run25c	113,368,438	112,216,242	80,597,194	62,110,624	29,917,240	2,115,118	1,534,146	3.05	72.53%	48.17%
		WQS115_wqs23.e1_hub05_L3_DN-NT.v8_100u_run25c	96,198,720	94,851,709	70,323,675	46,879,622	17,700,056	1,818,156	1,206,348	3.06	66.35%	37.76%
		WQS125_wqs30.e1_hub05_L3_DN-NT.v8_100u_run27c	75,823,374	75,015,149	53,716,257	43,616,620	34,829,300	3,381,646	2,527,992	4.37	74.76%	79.85%
DFF with end repair	N2 L3	WQS021_wqs02.e2_N2_L3_DFF-NT.v1_2u_run08c	51,516,956	42,439,388	36,773,473	24,569,022	18,054,394	895,342	443,608	2.9	49.55%	73.48%
DFF without end repair		WQS022_wqs02.e2_N2_L3_DFF-NT.v2_2u_run08c	13,559,754	12,469,858	10,496,569	7,336,074	4,160,706	253,042	83,792	3.1	33.11%	56.72%
biotin-11		WQS030_wqs02.e2_N2_L3_DN-NT-PD.v3_100u_run09c	22,389,468	21,811,727	18,489,571	12,148,862	177,950	26,504	19,516	2	73.63%	1.46%
biotin-14		WQS031_wqs02.e2_N2_L3_DN-NT-PD.v1_100u_run09c	2,831,504	2,747,237	2,548,775	1,427,530	33,534	6,394	4,548	3.5	71.13%	2.35%
biotin-adaptor		WQS029_wqs02.e2_N2_L3_DN-NT-PD.v2_50u_run09c	8,275,444	6,035,545	5,681,491	2,320,556	171,476	18,106	10,966	2.9	60.57%	7.39%
ngARC-C (A)		WQS032_wqs02.e2_N2_L3_MP-NT-PD.v1_2.5u_run09c	26,256,652	25,335,538	19,165,457	13,229,042	1,813,282	203,948	165,372	1.5	81.09%	13.71%
ngARC-C (B)		WQS033_wqs02.e2_N2_L3_MP-NT-PD.v2_2.5u_run09c	17,522,016	13,939,714	13,110,145	4,969,674	1,118,412	751,348	548,494	2	73.00%	22.50%
ngARC-C (C)		WQS034_wqs02.e2_N2_L3_MP-NT.v1_2.5u_run09c	34,661,904	33,818,559	29,706,242	20,748,632	17,128,934	426,728	323,658	2.1	75.85%	82.55%
ngARC-C (D)		WQS035_wqs02.e2_N2_L3_MP-TS-PD.v1_2.5u_run09c	2,361,950	2,160,719	2,143,562	524,646	398,566	163,062	96,802	2.7	59.37%	75.97%
ngARC-C (E)		WQS037_wqs02.e2_N2_L3_MP-NT-PD.v4_2.5u_run10c	53,922,148	46,770,662	30,207,995	16,362,982	15,310,150	248,646	182,762	1.5	73.50%	93.57%
ngARC-C (F)		WQS039_wqs02.e2_N2_L3_MP-NT-PD.v6_2.5u_run10c	142,868,866	142,484,983	142,293,671	108,450,634	102,081,770	1,111,058	1,088,700	1.4	97.99%	94.13%
ngARC-C (G)		WQS038_wqs02.e2_N2_L3_MP-NT-PD.v5_2.5u_run10c	54,634,944	50,083,124	48,830,263	30,724,422	7,834,542	644,096	484,928	1.6	75.29%	25.50%
ngARC-C (H)		WQS040_wqs02.e2_N2_L3_MP-TS-PD.v2_2.5u_run10c	7,777,300	6,761,394	6,318,436	2,165,100	654,648	248,944	203,884	1.3	81.90%	30.24%
ngARC-C (I)		WQS041_wqs02.e2_N2_L3_MP-TS-PD.v2_0.5u_run10c	4,319,012	3,409,283	3,151,296	771,396	144,026	42,162	33,742	1.5	80.03%	18.67%

Table A1.1: Summary of ARC-C library statistics.

ID	NameInThesis	SampleType	Culture condition	L2_1_%	L2_2_%	L3_3_%	L3_4_%	L3_5_%	L4_6_%
wqs01	N2_1a	N2 L3	Liquid		2.02	14.14	75.76	8.08	
wqs05	N2_2a/b		Liquid		3.09	12.37	79.39	4.12	
wqs33	N2_3		Liquid			9.76	79.27	10.97	
wqs14	*blmp-1*	*blmp-1* L3	Liquid		4.40	22.01	59.12	14.47	
wqs15			Liquid		5.06	15.19	64.56	15.19	
wqs16	*met-2 set-25*	*met-2 set-25* L3	Liquid			6.25	90.63	3.13	
wqs20			Liquid			4.41	91.57	4.02	
wqs73			Plate			3.16	95.43	1.41	
wqs30	hub02	hub02 L3	Liquid				88.89	11.11	
wqs39			Liquid				92.16	7.84	
wqs35	hub03	hub03 L3	Liquid			3.00	90.00	7.00	
wqs37			Liquid			2.08	94.79	3.13	
wqs23	hub05	hub05 L3	Liquid			12.31	80.00	6.73	0.96
wqs38			Liquid			8.53	84.53	6.93	
wqs65	MT13954	MT13954 L3	Plate			1.90	86.67	6.67	4.76
wqs72			Plate			1.06	92.55	6.38	
wqs63	MT16494	MT16494 L3	Plate			4.00	88.00	6.00	2.00
wqs67			Plate			4.04	92.93	3.03	
wqs17	MT17429	MT17429 L3	Plate		4.12	47.42	46.39	2.06	
wqs38			Plate		6.32	46.32	47.37		
wqs61	ST36	ST36 L3	Plate		14.16	76.99	7.96	0.88	
wqs71			Plate		11.11	78.70	10.19		
wqs57	N2_B	N2 L3	Plate	14.16	15.93	68.14	1.77		
wqs59	N2_D		Plate		16.48	41.76	39.56	2.20	
wqs64	N2_E		Plate		4.82	6.02	86.75	2.41	

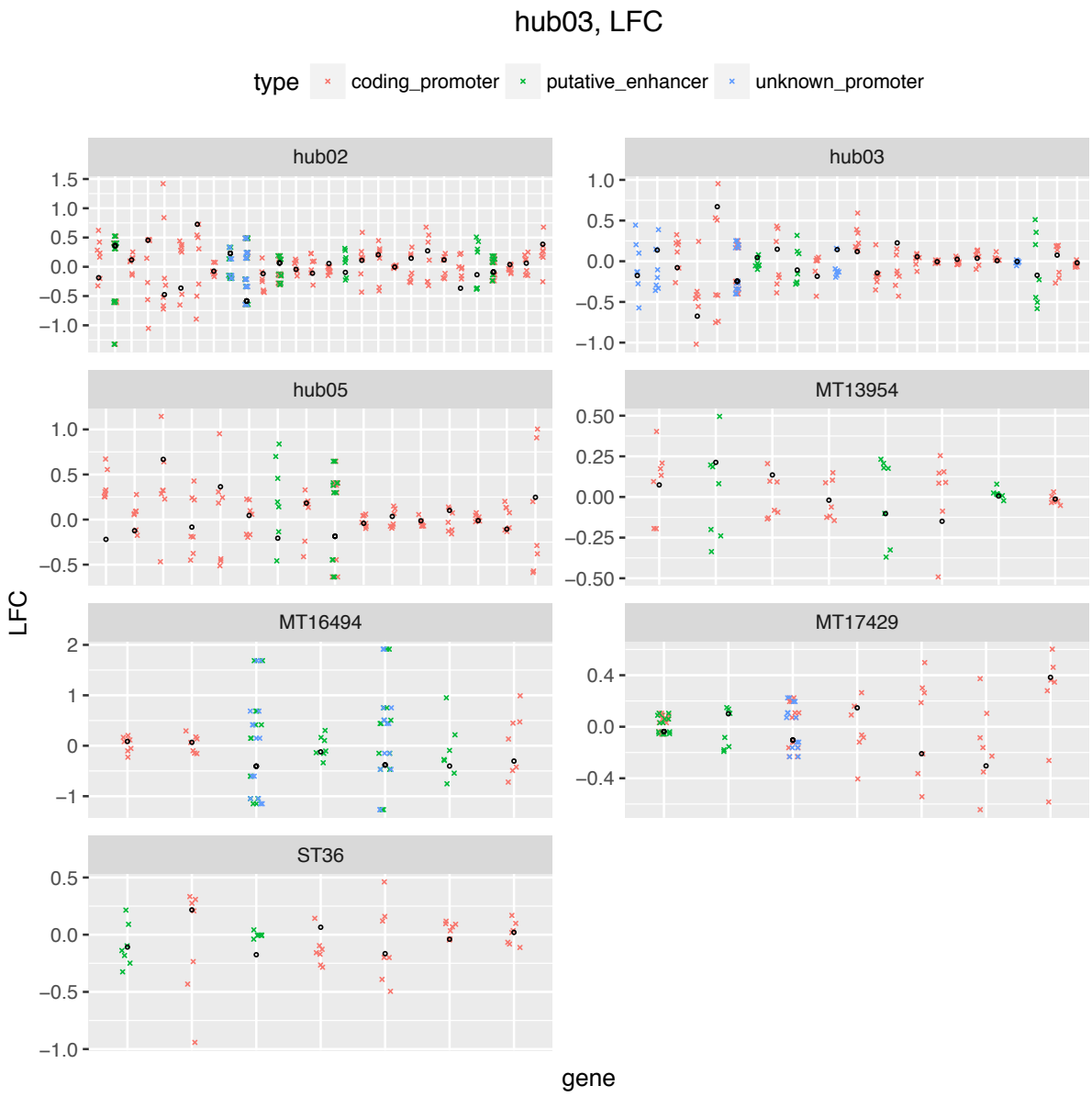
Table A1.2: Staging for worm collections.

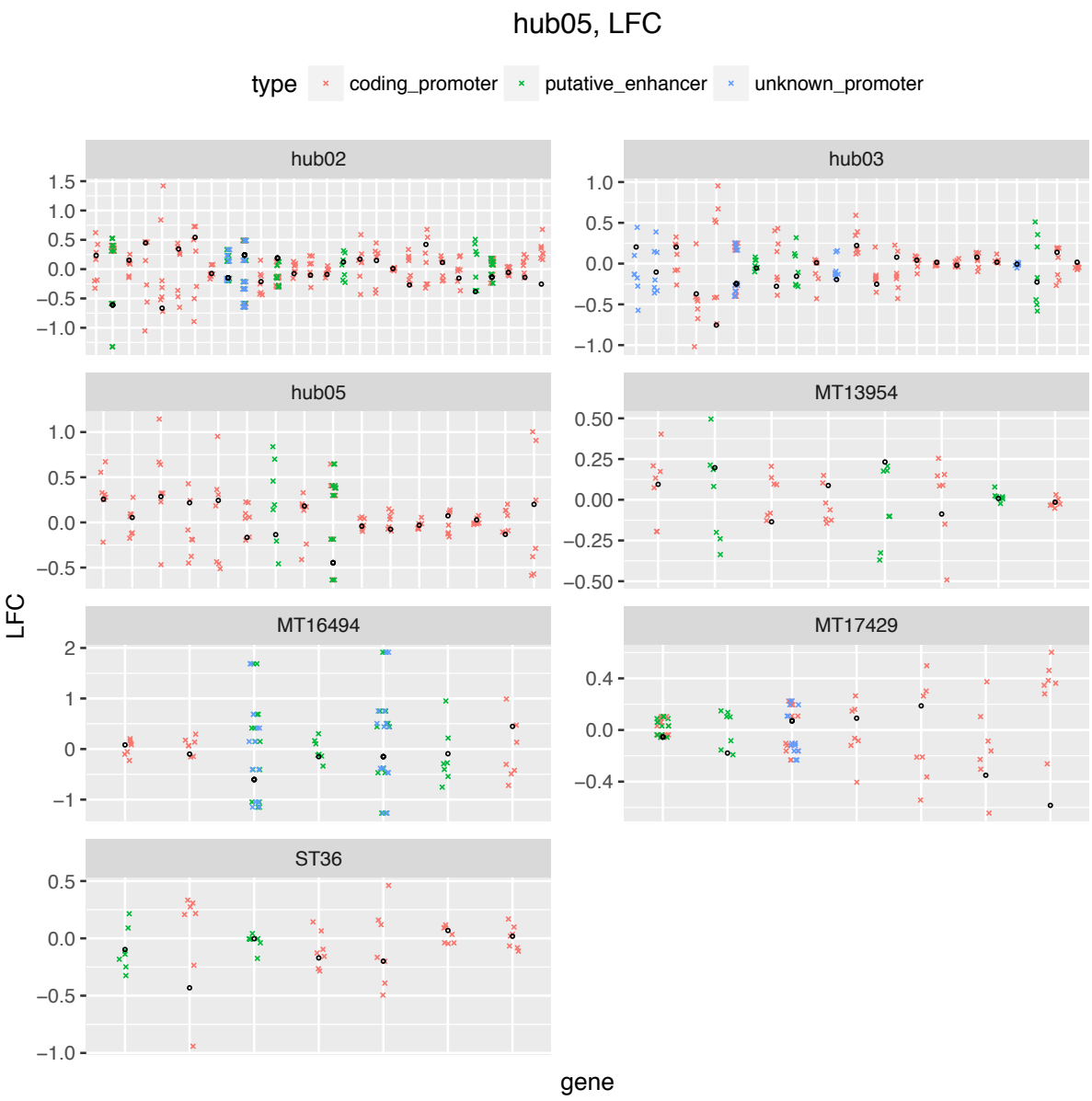
Strain	Strain comparison - p value
hub02	0.5487817
hub03	0.8786290
hub05	0.9642082
MT13954	0.7265183
MT16494	0.4637841
MT17429	0.8535111
ST36	0.3436966

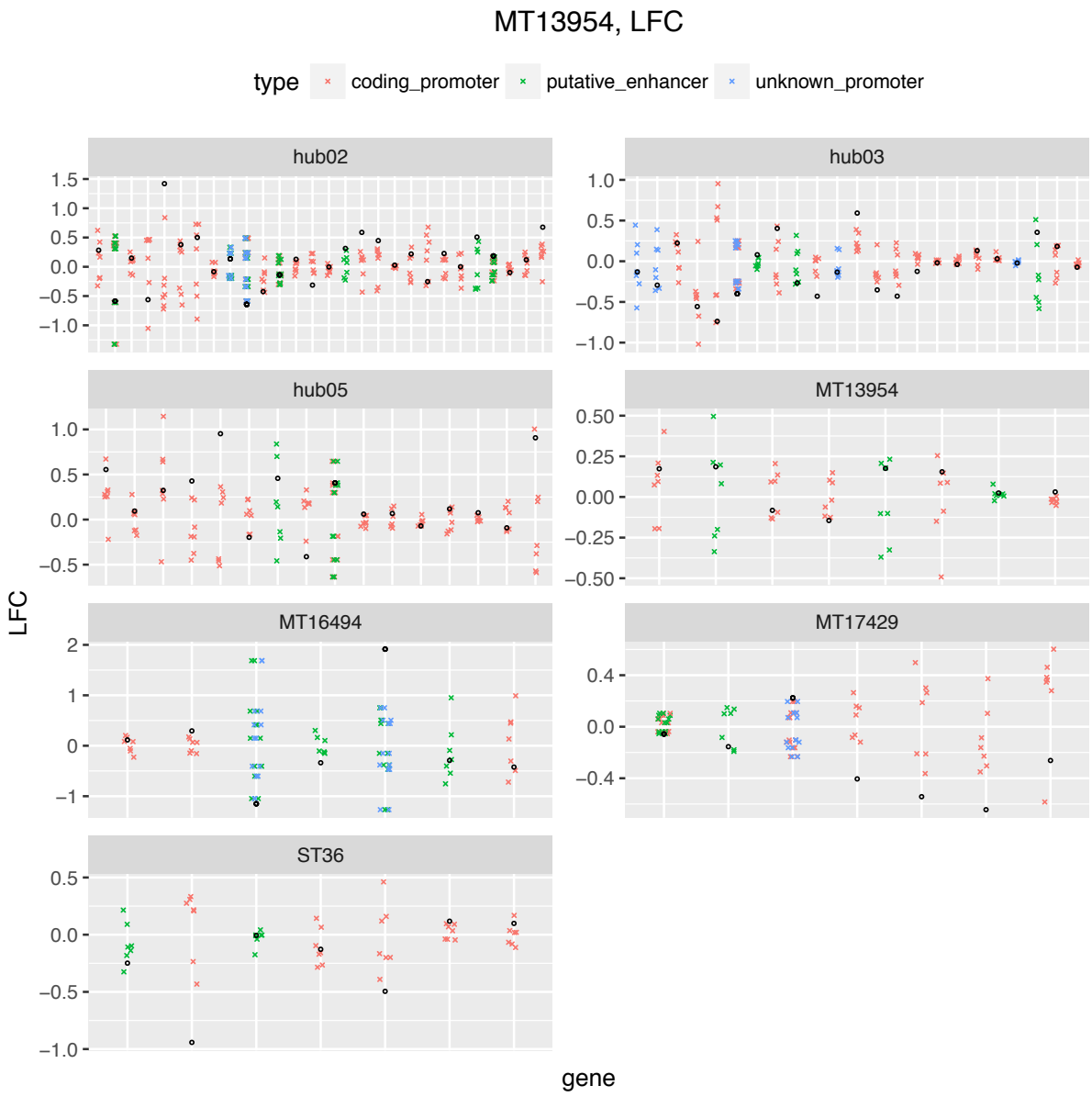
Table A1.3: P-values for linked-gene analysis. Fold-changes of gene expression in linked genes in the strain of interest were compared with those from linked genes in other strains (one-sided t-test).

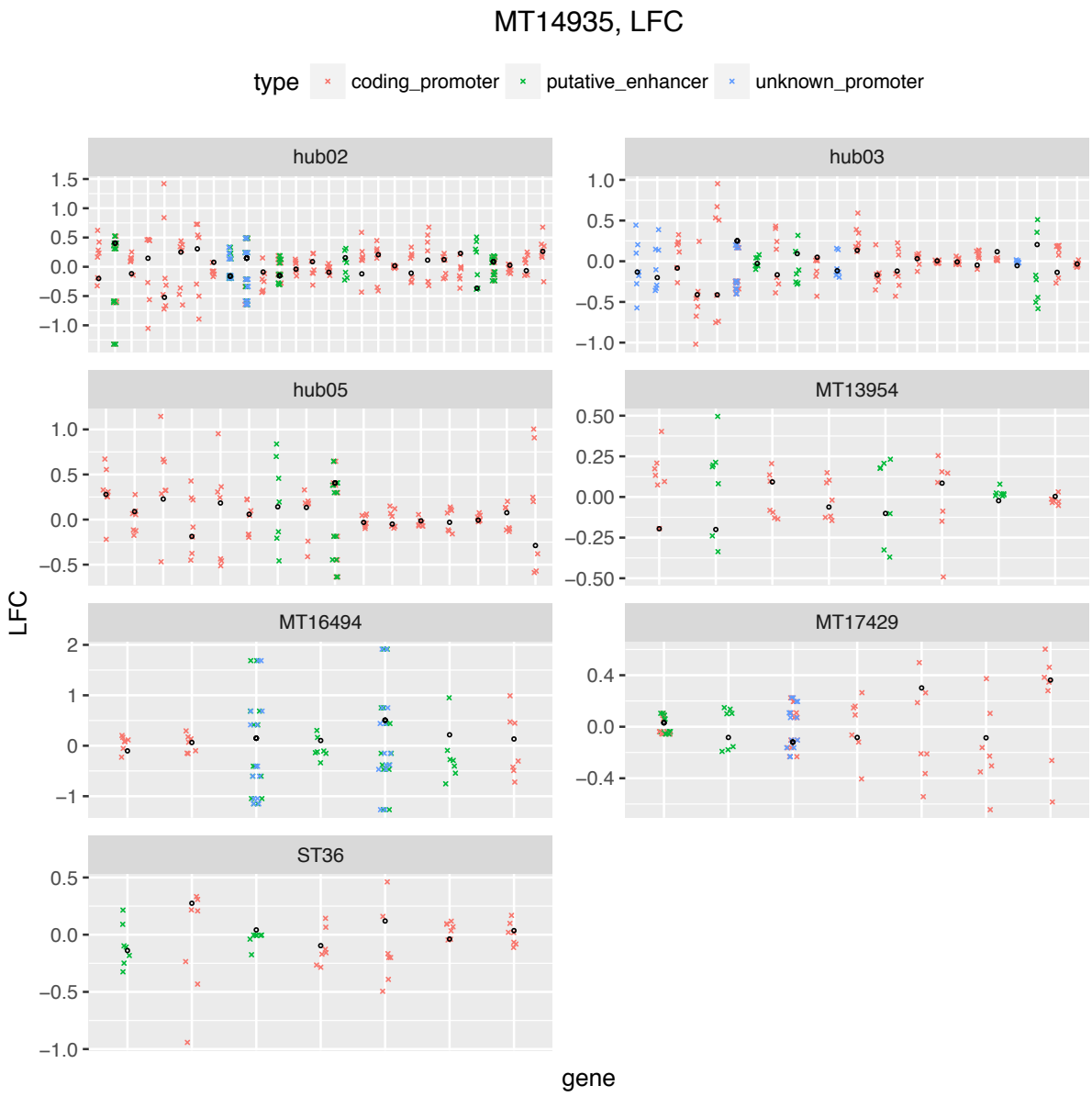
strain	FC.compare.window.p	max.d
hub02	0.39835204	1e+04
hub02	0.12139351	5e+04
hub02	0.57479063	1e+05
hub02	0.55838908	2e+05
hub02	0.99874542	1e+06
hub03	0.74735006	1e+04
hub03	0.94382197	5e+04
hub03	0.98897797	1e+05
hub03	0.86383982	2e+05
hub03	0.77863151	1e+06
hub05	NA	1e+04
hub05	0.54424668	5e+04
hub05	0.96259519	1e+05
hub05	0.77029818	2e+05
hub05	0.02377153	1e+06
MT13954	NA	1e+04
MT13954	0.91650836	5e+04
MT13954	0.93854814	1e+05
MT13954	0.97176543	2e+05
MT13954	0.99855460	1e+06
MT16494	NA	1e+04
MT16494	0.45837263	5e+04
MT16494	0.86092811	1e+05
MT16494	0.98547838	2e+05
MT16494	0.24678267	1e+06
MT17429	0.64644258	1e+04
MT17429	0.68385677	5e+04
MT17429	0.02468729	1e+05
MT17429	0.35328306	2e+05
MT17429	0.97296844	1e+06
ST36	NA	1e+04
ST36	0.86874071	5e+04
ST36	0.96931235	1e+05
ST36	0.93976216	2e+05
ST36	0.74255061	1e+06

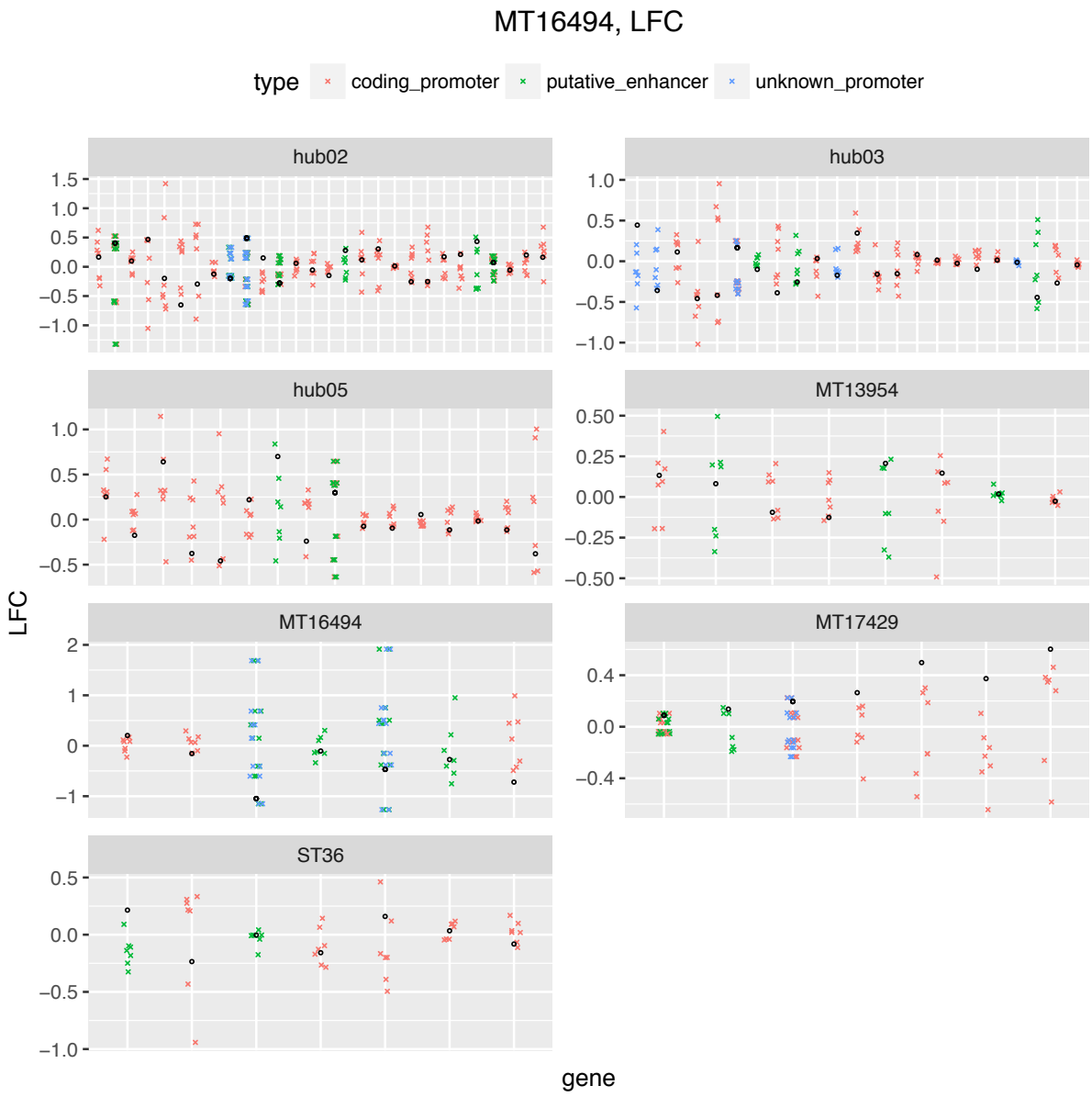
Table A1.4: P-values for local-genes analysis. Fold-changes of genes within different distance intervals (max.d) centred on deletion of interest were compared with genes in other regions of the genome of similar distance intervals.


hub03, LFC
type
coding_promoter
putative_enhancer
unknown_promoter
hub02
hub03
hub05
MT13954
MT16494
MT17429
ST36
LFC
gene


hub05, LFC
type
coding_promoter
putative_enhancer
unknown_promoter
hub02
hub03
hub05
MT13954
MT16494
MT17429
ST36
LFC
gene


MT13954, LFC
type
coding_promoter
putative_enhancer
unknown_promoter
hub02
hub03
hub05
MT13954
MT16494
MT17429
ST36
LFC
gene


MT14935, LFC
type
coding_promoter
putative_enhancer
unknown_promoter
hub02
hub03
hub05
MT13954
MT16494
MT17429
ST36
LFC
gene


MT16494, LFC
type
coding_promoter
putative_enhancer
unknown_promoter
hub02
hub03
hub05
MT13954
MT16494
MT17429
ST36
LFC
gene

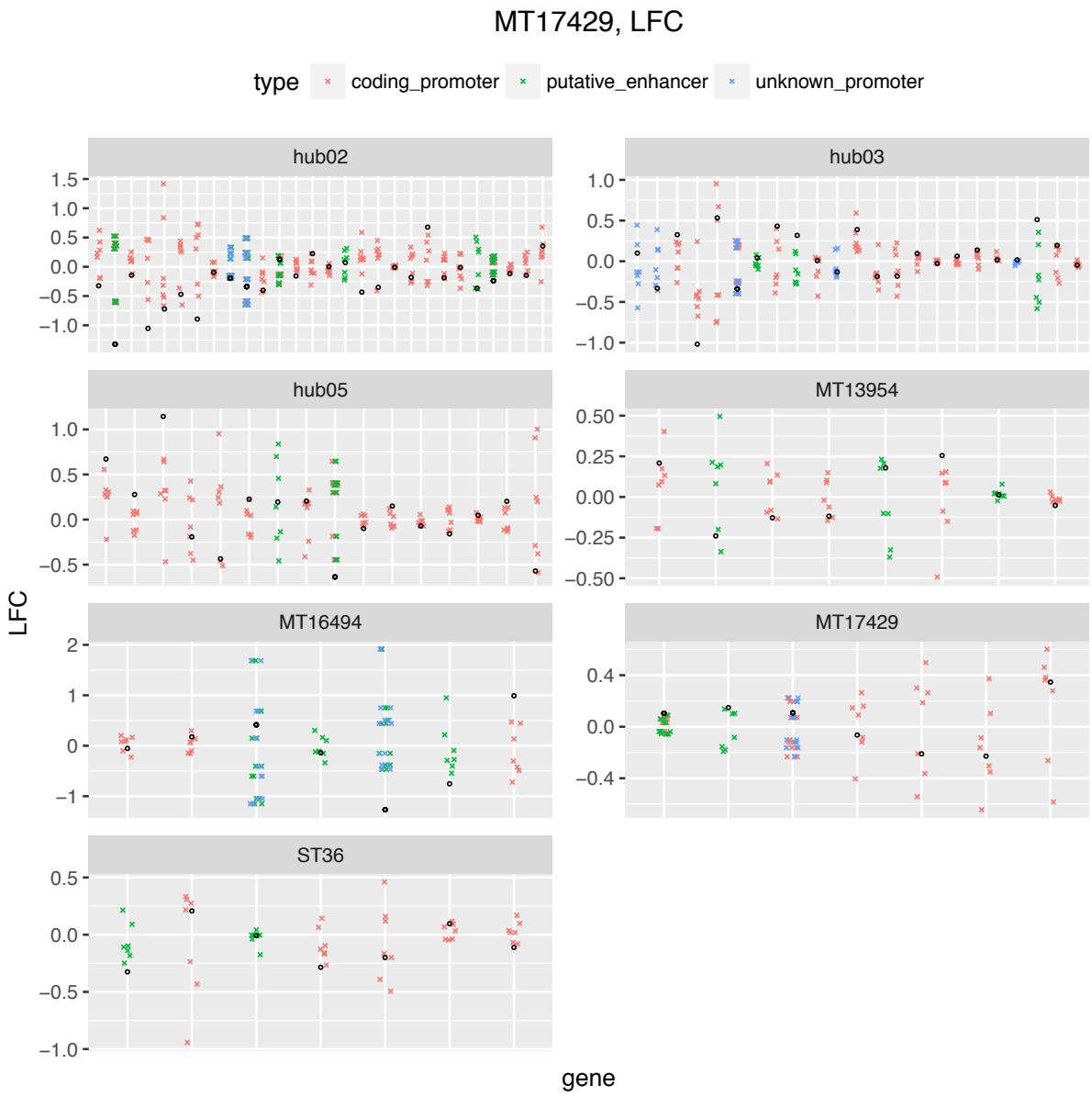
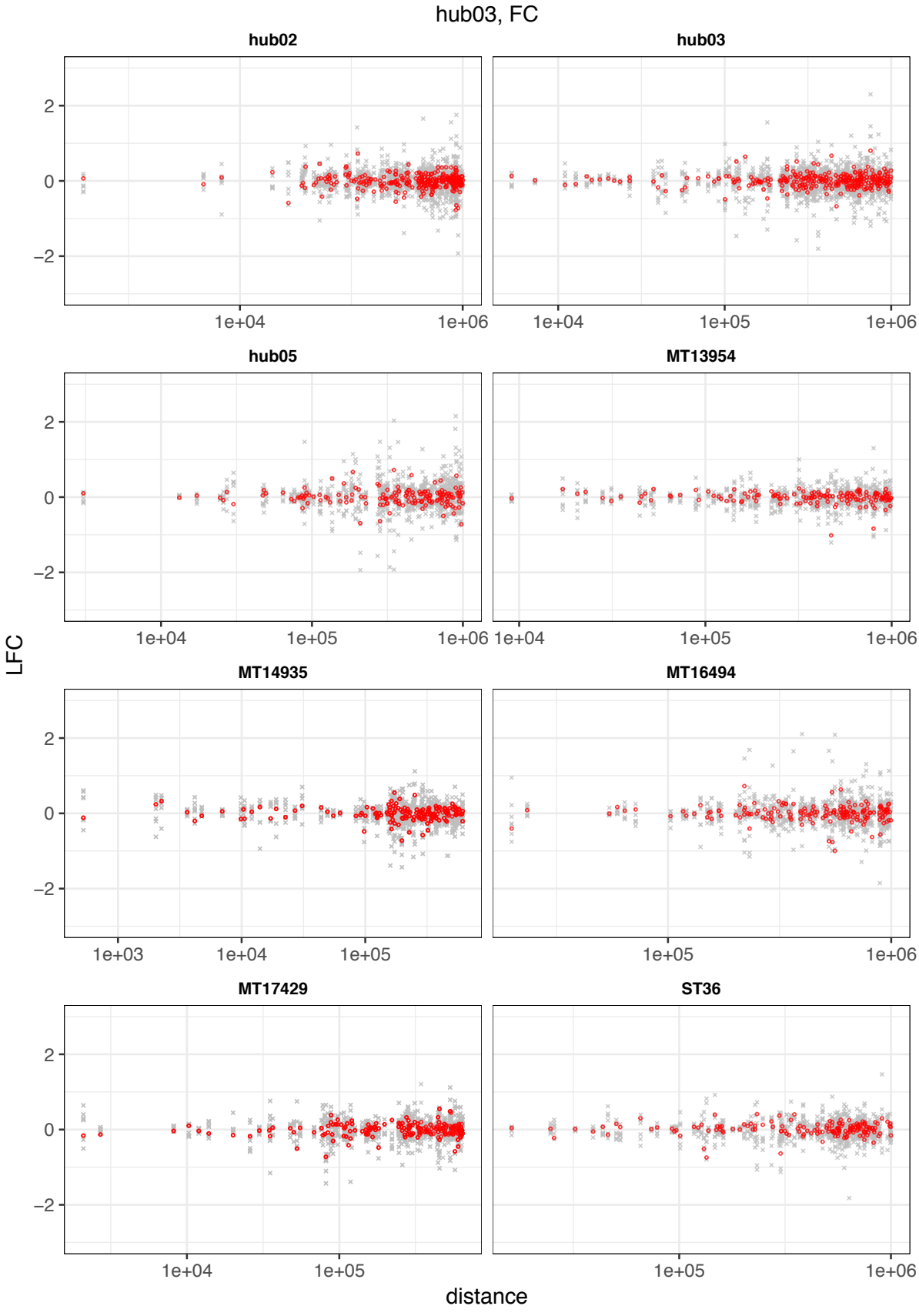
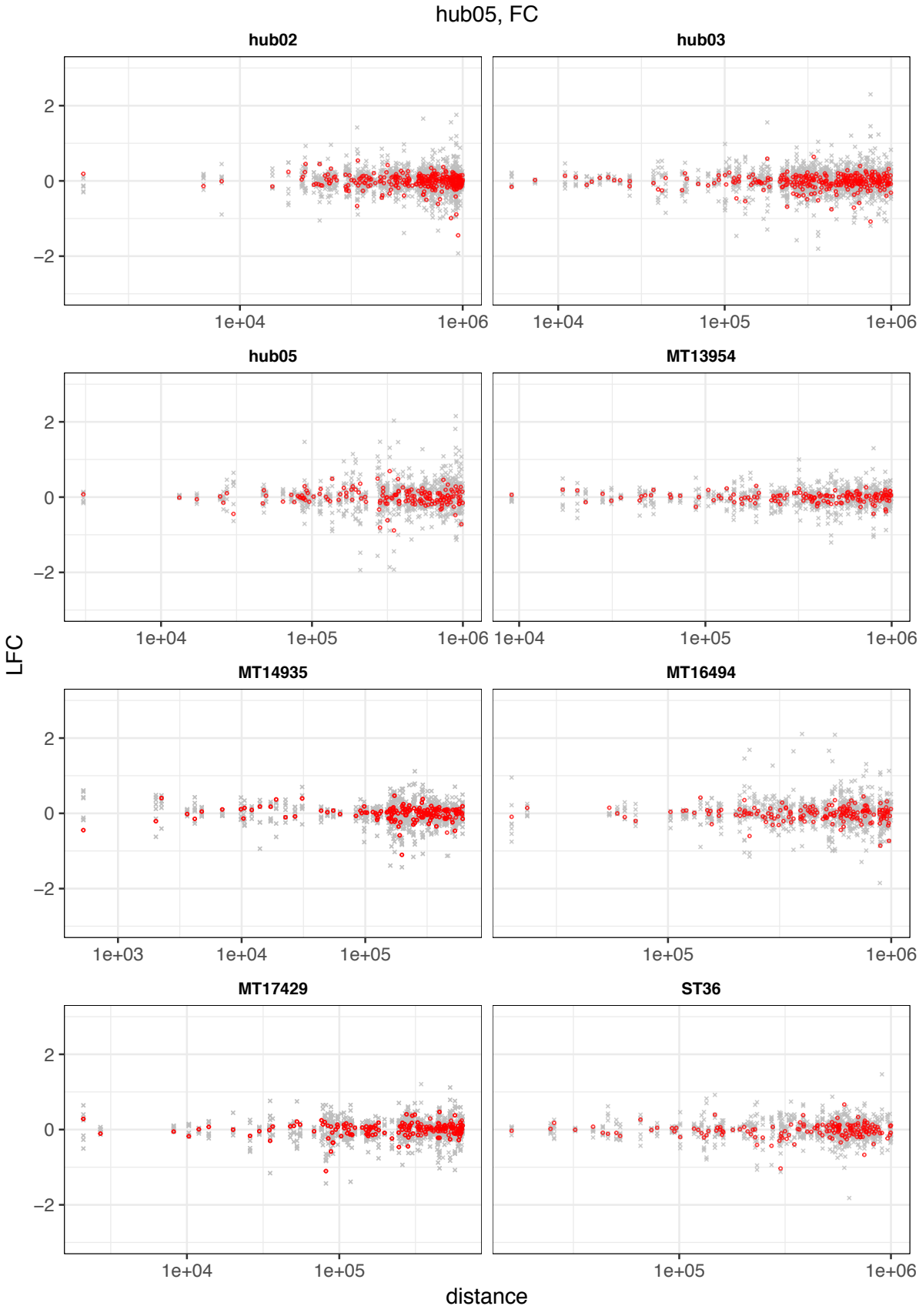
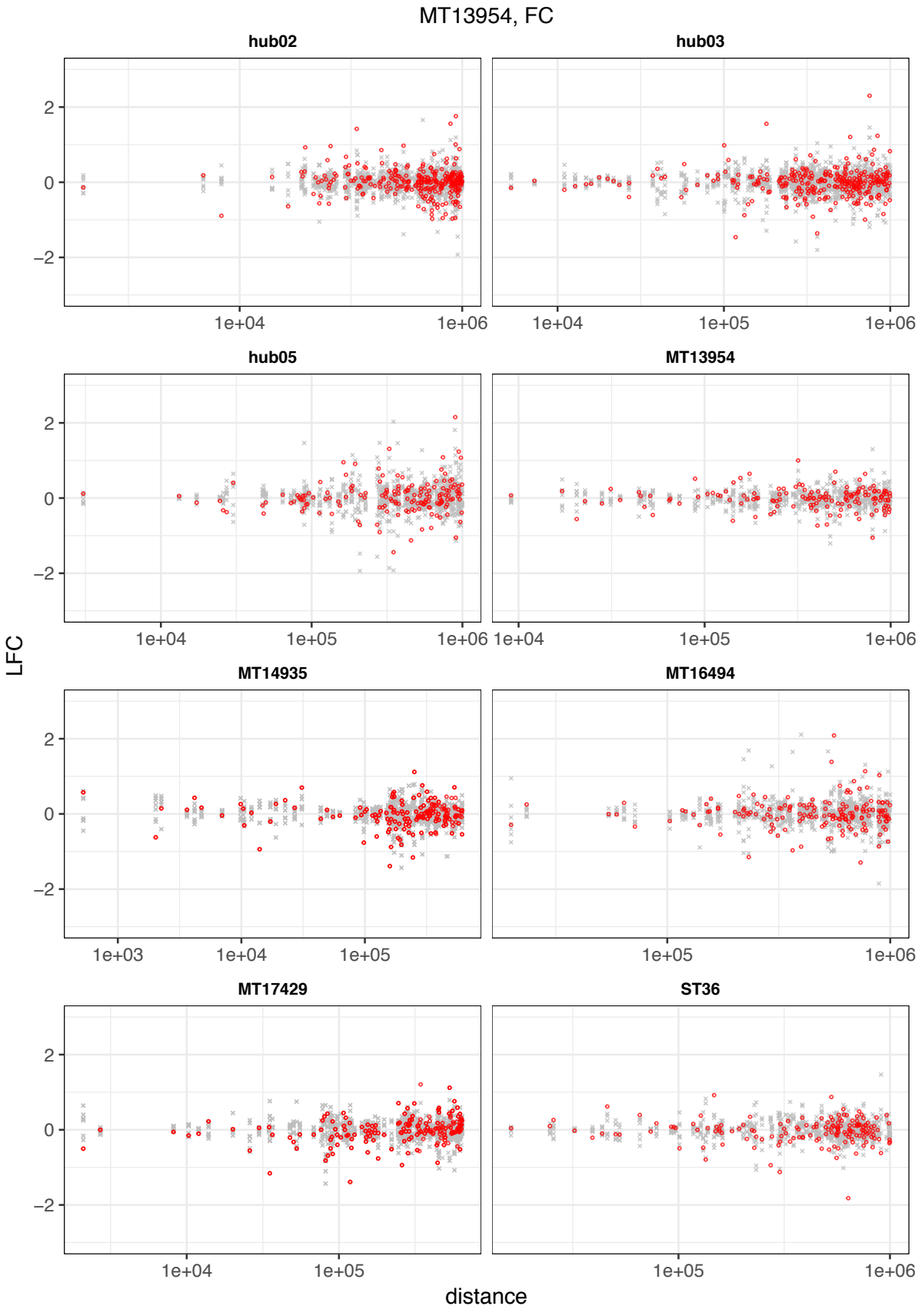
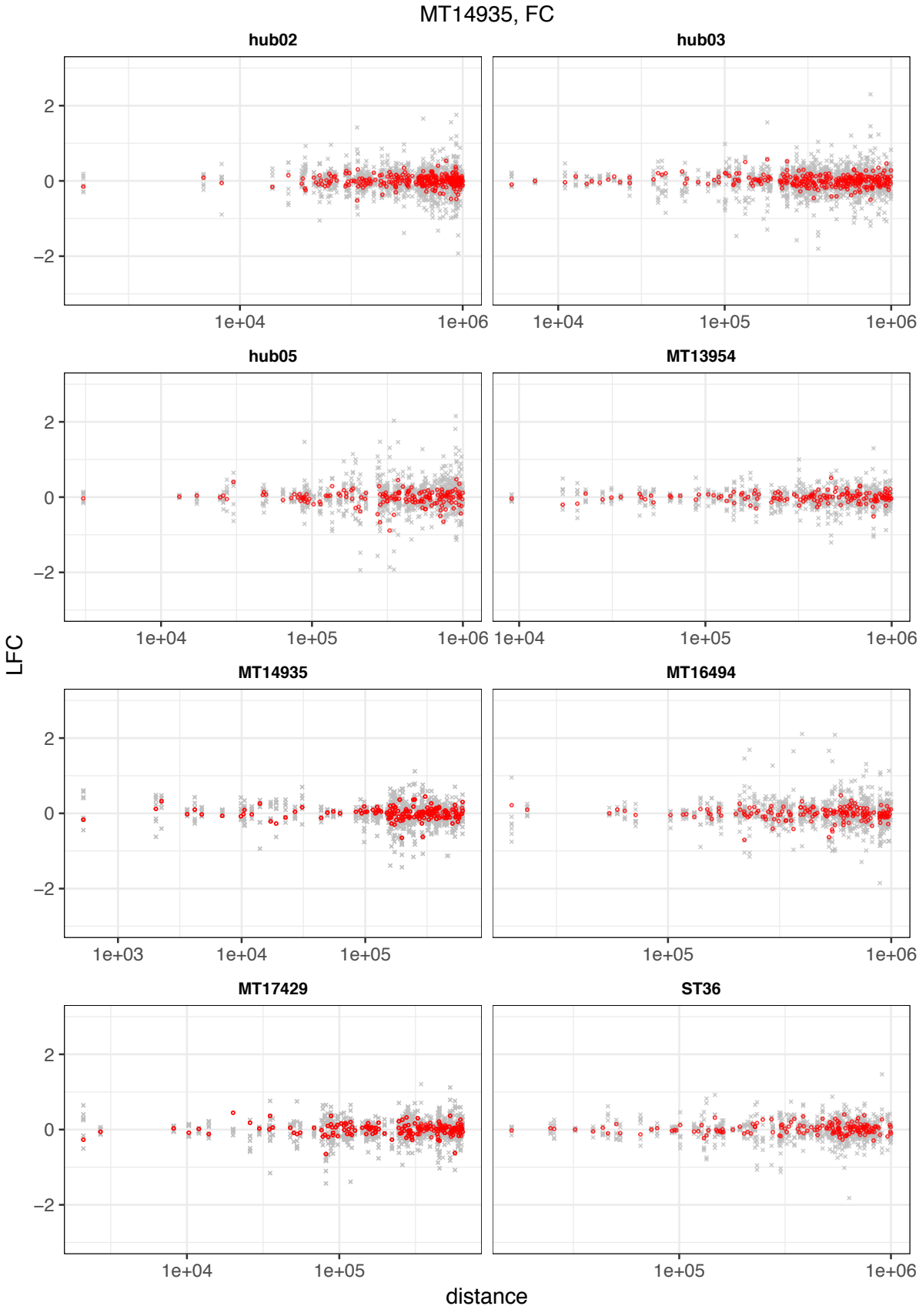

MT17429, LFC
type
coding_promoter
putative_enhancer
unknown_promoter
hub02
hub03
hub05
MT13954
MT16494
MT17429
ST36
LFC
gene

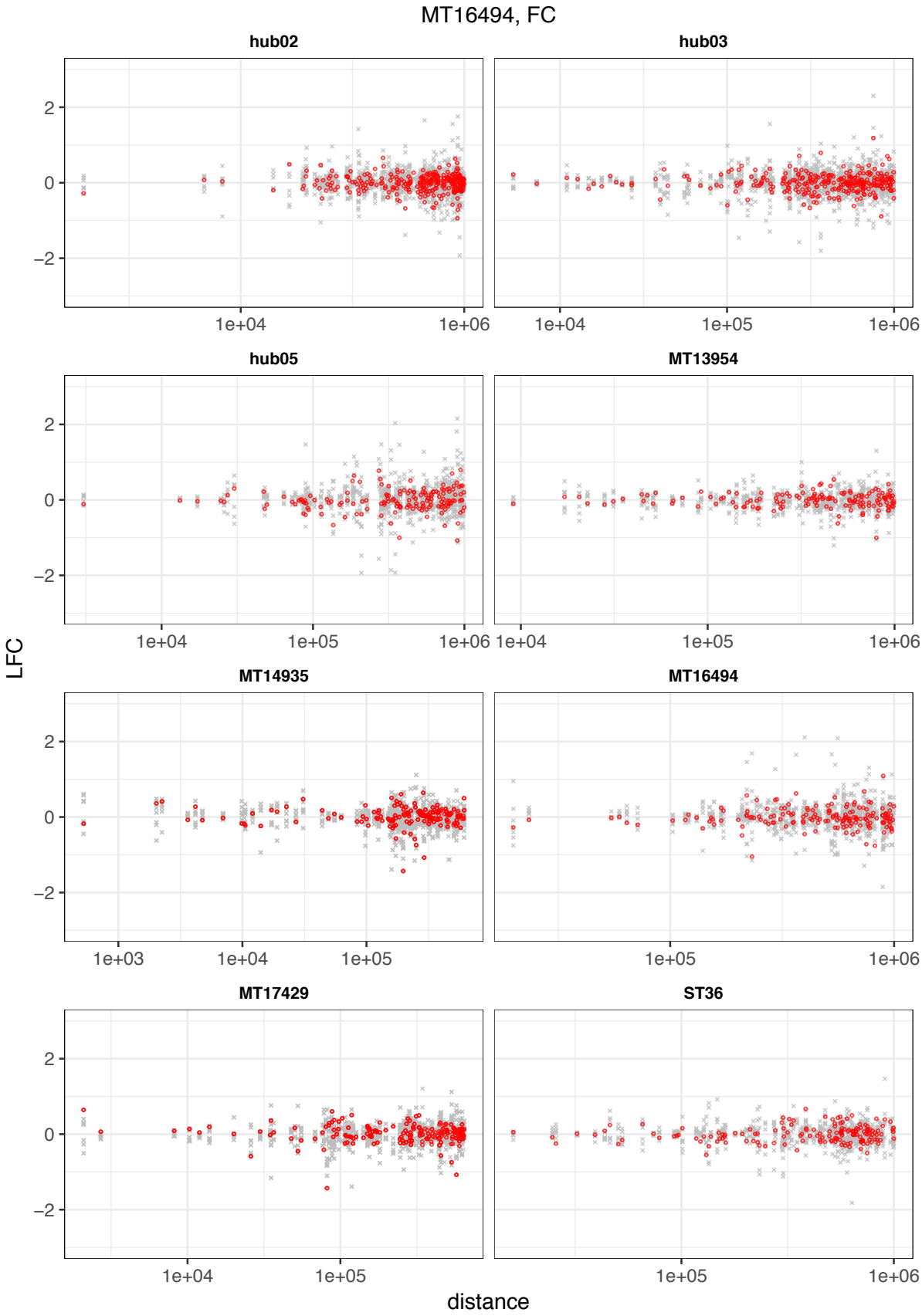
Figure A1.5: Linked-genes analysis for hub deletions. Expression lg2FC was calculated for each deletion's own linked genes and compared with linked genes from other strains. Black circles are lg2FCs in the strain of interest, while coloured crosses are lg2FCs of the same genes in other strains.

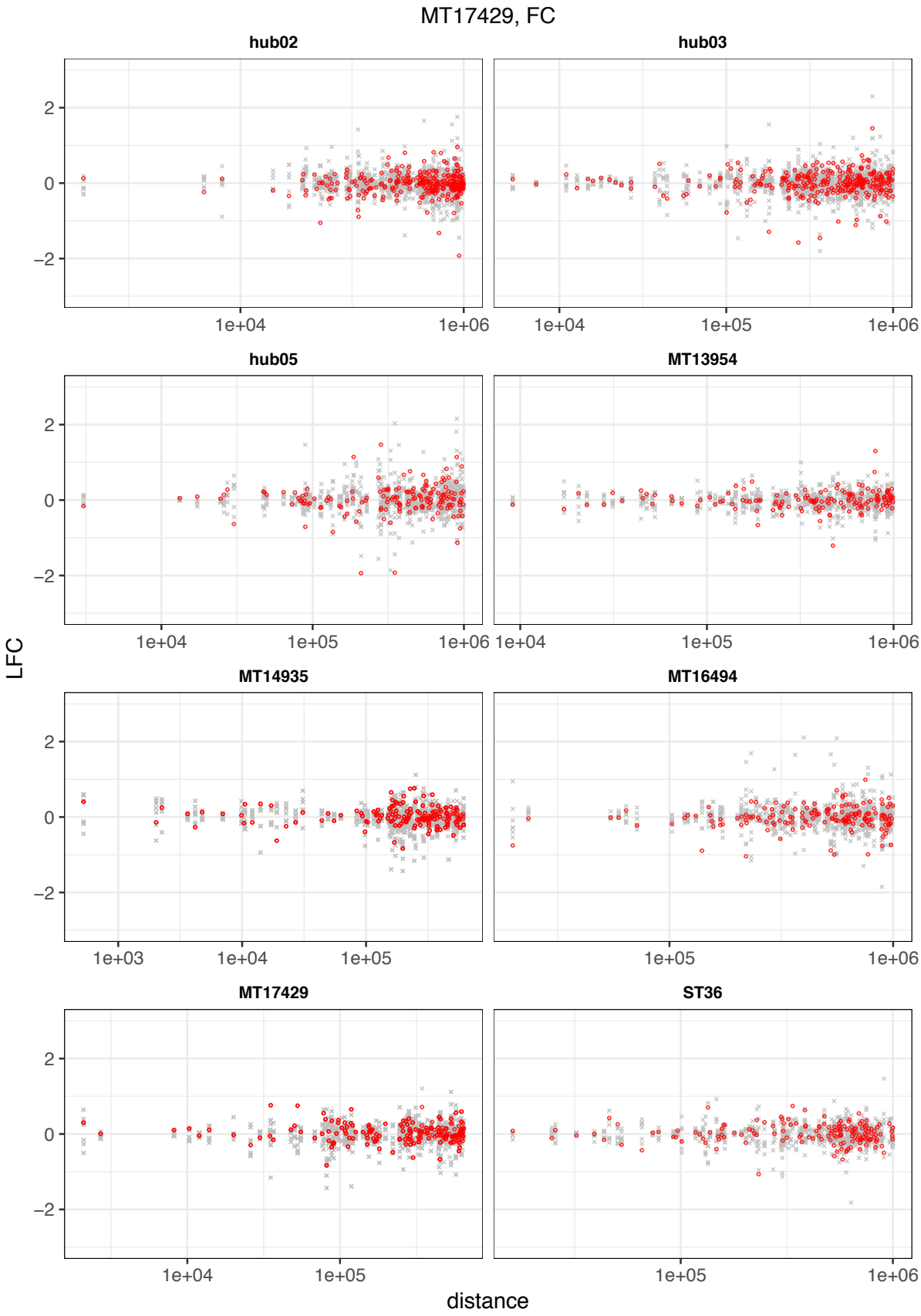

hub03, FC
hub02
hub03
hub05
MT13954
MT14935
MT16494
MT17429
ST36
LFC
distance
2
0
−2
1e+03
1e+04
1e+05
1e+06


hub05, FC
hub02
hub03
hub05
MT13954
MT14935
MT16494
MT17429
ST36
LFC
distance
2
0
−2
1e+03
1e+04
1e+05
1e+06


MT13954, FC
hub02
hub03
hub05
MT13954
MT14935
MT16494
MT17429
ST36
LFC
distance
2
0
−2
1e+03
1e+04
1e+05
1e+06


MT14935, FC
hub02
hub03
hub05
MT13954
MT14935
MT16494
MT17429
ST36
LFC
distance
2
0
−2
1e+03
1e+04
1e+05
1e+06


MT16494, FC
hub02
hub03
hub05
MT13954
MT14935
MT16494
MT17429
ST36
LFC
distance
2
0
−2
1e+03
1e+04
1e+05
1e+06


MT17429, FC
hub02
hub03
hub05
MT13954
MT14935
MT16494
MT17429
ST36
LFC
distance
2
0
−2
1e+03
1e+04
1e+05
1e+06

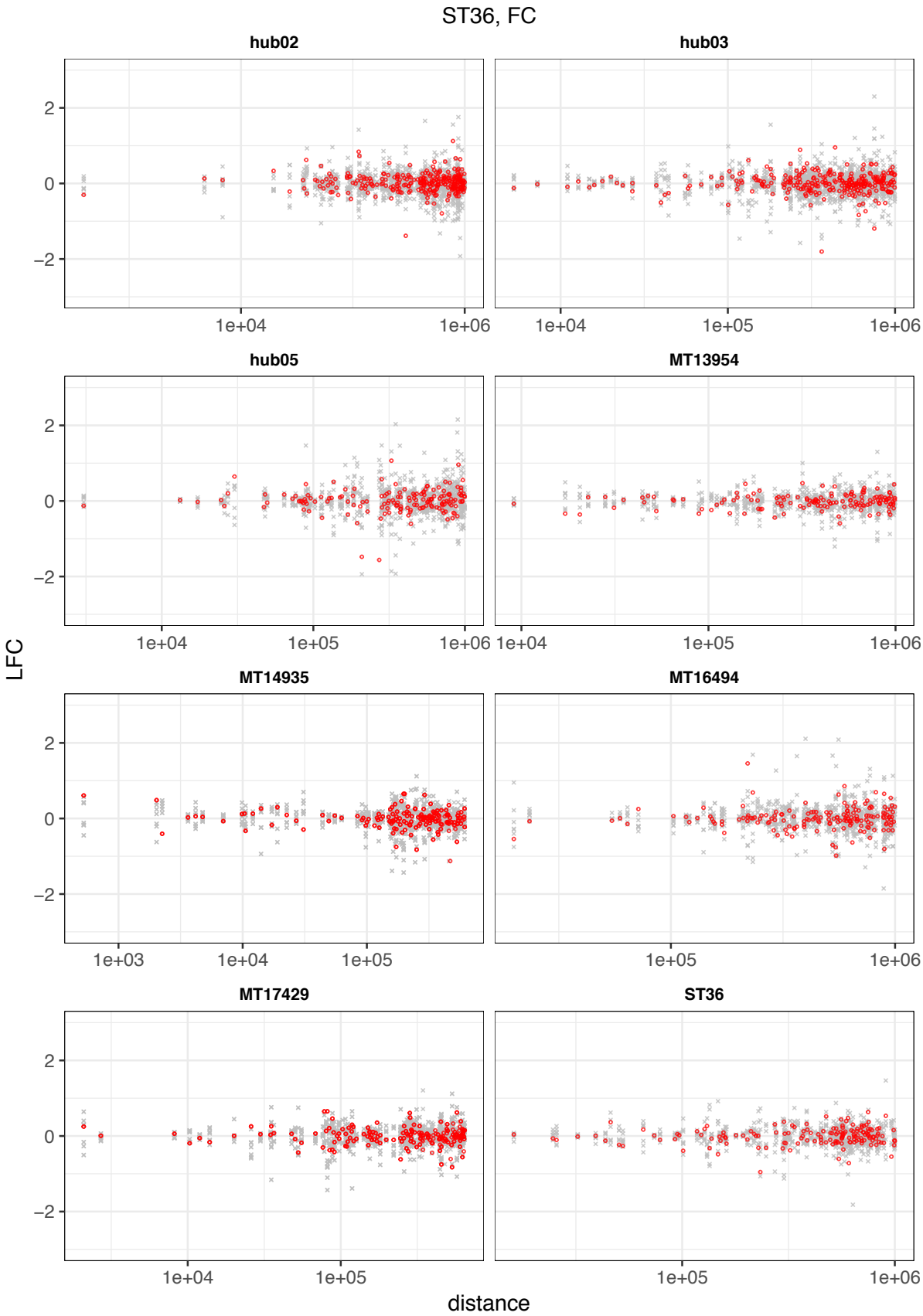


Figure A1.6: Local genes analysis for hub deletions. Expression lg2FC of genes is plotted as a function of genomic distance from the deletion site. Red circles represent lg2FC of genes in the deletion strain of interest, grey crosses represent

lg2FC of the same genes in other strains.